Estimating Impact with Surveys versus Digital Traces: Evidence from Randomized Cash Transfers in Togo*

Emily Aiken Suzanne Bellue Joshua E. Blumenstock
Dean Karlan Christopher Udry

September 23, 2023

Abstract

Do non-traditional digital trace data and traditional survey data yield similar estimates of the impact of a cash transfer program? In a randomized controlled trial of Togo's COVID-19 Novissi program, endline survey data indicate positive treatment effects on beneficiary food security, mental health, and self-perceived economic status. However, impact estimates based on mobile phone data – processed with machine learning to predict beneficiary welfare – do *not* yield similar results, even though related data and methods do accurately predict wealth and consumption in prior cross-sectional analysis in Togo. This limitation likely arises from the underlying difficulty of using mobile phone data to predict short-term changes in wellbeing within a rural population with fairly homogeneous baseline levels of poverty. We discuss the implications of these results for using new digital data sources in impact evaluation.

JEL Codes: C55, I32, I38

*Aiken: emilyaiken@berkeley.edu, University of California Berkeley; Bellue: sbellue@mail.uni-mannheim.de, CREST - ENSAE; Blumenstock: jblumenstock@berkeley.edu, University of California Berkeley; Karlan: karlan@kellogg.northwestern.edu, Northwestern University; Udry: christopher.udry@northwestern.edu, Northwestern University. Shikhar Mehra and Nathaniel Ver Steeg provided excellent research assistance and Isabel Oñate provided excellent research management. This project was possible through the dedication of our project partners in Togo, especially C. Lawson, S. Bakari, L. Mills, K. Ekouhoho, M. Koudeka and A. Byll. We are grateful for support from Institut National de la Statistique et des Études Economiques et Démographiques (INSEED), led by S. Telou and R. Ogoumedi; as well as the team at GiveDirectly, especially H. S. Chia, M. Cooke, K. Lee, A. Nawar and D. Quinn. Simón Ramirez Amaya and Ethan Ligon provided helpful feedback. We thank the World Bank for supporting the phone surveys and data collection under the WURI program. Aiken gratefully acknowledges financial support from a Microsoft Research PhD Fellowship. Bellue gratefully acknowledges financial support from the German Academic Exchange Service (DAAD) and the German Research Foundation (through the CRC-TR-224 project A03). Blumenstock thanks the National Science Foundation for support under CAREER Grant IIS-1942702. This trial was pre-registered with the American Economic Association Registry #7590. IRB approval was obtained from the U.C. Berkeley Committee for the Protection of Human Subjects (Protocol #2020-05-13281).

1 Introduction

Reliable estimates of post-program outcomes are essential to impact evaluation. In low- and middle-income countries (LMICs), such outcomes are traditionally measured through surveys. However, a new paradigm is emerging for estimating living standards in LMICs, based on the application of machine learning algorithms to nontraditional data from mobile phones (Blumenstock et al., 2015; Blumenstock, 2018), satellites (Jean et al., 2016; Chi et al., 2022), and other digital sources (Sheehan et al., 2019; Fatehkia et al., 2020). These estimates are attractive because they can be produced rapidly for large populations at a fraction of the cost of traditional surveys. Policymakers are also increasingly using such data sources for targeting: for instance, the governments of Togo and the Democratic Republic of Congo both used phone-based wealth estimates to identify cash transfer recipients (Aiken et al., 2022b; Mukerjee et al., 2023); the government of Colombia incorporated digital financial records in an aid targeting algorithm (Lopez, 2021); and the government of Nigeria used satellite-based poverty maps to select urban areas for government subsidies (Smythe and Blumenstock, 2022).

We ask whether welfare outcomes estimated from "digital trace" data produce the same estimates of program impact as those obtained from traditional survey-based measures of welfare. We study these questions in the context of Togo's *Novissi* program, which provided five monthly cash transfers of roughly USD $14 to poor individuals in rural Togo during the COVID-19 pandemic. The program was rolled out in two phases (starting either in late 2020 or the middle of 2021), and through a randomized controlled trial (RCT) individuals were randomly assigned to receive transfers during one of the two phases. We conducted phone surveys in between the phases, and also obtained the complete mobile phone transaction logs of all consenting program participants.

Our first set of results uses the phone surveys to document the welfare impacts of the *Novissi* program, using traditional methods prespecified in a pre-analysis plan. We find that Novissi transfers significantly increased food security (by 0.06 standard deviations (SD), standard error (se)=0.02), mental health (0.07 SD, se=0.02), and self-perceived economic status (0.04 SD, se=0.02) of transfer recipients. Effects on other welfare outcomes are positive but not statistically significant. The total program effect on a composite index of welfare is 0.06 SD (se=0.02).

Our second set of results relies on post-program estimates of welfare derived from the mobile phone data of beneficiaries and non-beneficiaries. Machine learning algorithms applied to the phone data can predict a proxy means test (PMT) measure of poverty with accu-

racy comparable to that seen in prior work (Blumenstock et al., 2015; Blumenstock, 2018; Aiken et al., 2022b), but they are much less effective at estimating the measures of welfare that were impacted by the program (i.e., food security, mental health, and self-perceived economic status). We then use these predicted measures of welfare to estimate program impacts, and find that the estimates based on phone records differ considerably from those based on surveys.

After presenting the main results, we conduct analysis to understand why food security and other measures of vulnerability could not be accurately predicted from phone data, and how that contributes to the inconsistency between program impacts estimated from phone surveys versus mobile phone metadata. These results highlight several challenges associated with estimating vulnerability measures from the phone data of a very homogenous population. In particular, the fact that poverty in Togo is more geographically concentrated than vulnerability helps explain why vulnerability is quite difficult to predict from mobile phone data. We also show that – even if the predictive models had been more accurate – impact evaluation using phone data could still be complicated by issues of model drift, and by the difficulty of inferring impacts that were modest in magnitude. We conclude with a discussion of how these results can inform the broader conversation around the use of digital data for monitoring and impact evaluation.

1.1 Related Work

We primarily contribute to two main literatures: first, the literature on impacts of unconditional cash transfers in emergency contexts (and more specifically, in the context of the COVID-19 pandemic), and second, the literature on using digital data sources for measuring welfare and for targeting and evaluating programs and policies.

Cash Transfer Impacts in Humanitarian Contexts

Many studies, both experimental and quasi-experimental, have documented impacts of unconditional cash transfers on a range of welfare outcomes, including household expenditures, food security, health, education, savings, and financial inclusion (for a review, see Bastagli et al. (2016)). Prior to the COVID-19 pandemic, there was limited evidence on the impact of cash transfers distributed in response to crises; exceptions include transfers during droughts in Niger (Aker et al., 2011), to Colombian refugees in Ecuador (Hidrobo et al., 2014), and to conflict-affected households during Yemen's civil war (Schwab, 2013) — Doocy and Tappis (2017) reviews this sub-literature on transfers in humanitarian emergencies. Since the

COVID-19 pandemic, a growing body of research has emerged to document the welfare impacts of cash transfers distributed in response to the pandemic. Many of these studies are reviewed in Karlan et al. (2022). Broadly, this literature shows modest, positive, and statistically significant impacts of cash transfers on a wide range of welfare metrics, including food security and mental health.[1]

The first portion of our analysis contributes to this literature by documenting the impacts of pandemic cash transfers in Togo, using a large cash transfer program where treatment was randomly assigned at the individual level. While the cash transfers we study are smaller ($13-15.50 per month) than most of the other programs studied ($15-50 per month), we document comparable effect sizes (0.04-0.07 standard deviations).

Digital Data for Targeting and Evaluating Social Protection Programs

Our second major focus — to test whether the impact of cash transfers can be estimated with mobile phone data — builds on a line of work estimating welfare from digital data. Following Blumenstock (2014), a handful of papers document the ability of mobile phone records, combined with machine learning (ML) techniques, to estimate the wealth and consumption of individual mobile phone subscribers in Rwanda (Blumenstock et al., 2015), Afghanistan (Blumenstock, 2018; Aiken et al., 2022a), Togo (Aiken et al., 2022b), and an undisclosed low-income Asian country (Sundsøy et al., 2016). While this literature shows that phone data can be used to estimate levels of poverty — and how those estimates can be used to determine eligibility for cash transfers (Aiken et al., 2022b) — it does not ask whether such measurements can be used to infer the impact of cash transfers or other policies or interventions.[2] We help fill this gap by providing new evidence on the potential — and

[1] Banerjee et al. (2020) use phone surveys and an RCT design to show that universal basic income transfers of USD 22.5 nominal per month to households under lockdown in Kenya reduced the probability of households experiencing hunger (by 5-11 percentage points, relative to a control mean of 68%), and had modest positive impacts on mental health. Similarly, Londoño-Vélez and Querubin (2022) use an RCT and phone surveys to measure impacts of a monthly VAT refund of USD 19 in Colombia, finding a 4.4 percentage point increase in the probability of treated households purchasing food in the week preceding the survey (relative to a control mean of 72%), but no statistically significant impacts on food security. The paper also reports positive and statistically significant impacts on mental health indices (1.2-2.1 percentage points) and a financial health index (0.055 standard deviations). Karlan et al. (2022) follow a similar experimental design, using an RCT and several rounds of phone surveys to evaluate the impact of eight monthly cash transfers of $15, recording an 8% increase in food consumption among treated households. In a non-randomized approach, Bottan et al. (2021) use online surveys and a regression discontinuity design to show that pension payments of USD 43-50 per month in Bolivia decreased the probability of households going hungry by 8-12 percentage points, relative to a comparison mean of 22%.

[2] One exception is the ongoing work by Barriga-Cabanillas et al. (nd), which uses a regression discontinuity design to study the impact of a cash transfer program in Haiti. Preliminary results, consistent with our own, suggest that phone-based estimates of food security are too noisy to detect the impact of cash transfers.

limitations — for using mobile phone data and machine learning for impact evaluation.

Our work also connects to recent efforts to use remote sensing data for impact evaluation. Following a number of papers showing that wealth can be estimated from satellite imagery (Jean et al., 2016; Yeh et al., 2020; Chi et al., 2022), two recent papers estimate the impacts of development interventions from satellite imagery. Huang et al. (2021) use satellite and ML-derived estimates of housing quality to estimate the long-run impacts of large, one-time cash transfers (USD 1,000) in rural Kenya; they find that impact estimates based on imagery are similar to, but noisier than, estimates based on surveys. Ratledge et al. (2021) use satellite and ML-derived estimates of wealth to evaluate the impact of electrical grid expansion in Uganda, finding point estimates in a similar range to a difference-in-differences strategy exploiting only household survey data.

2 The GiveDirectly-Novissi Program: Design and Data

The GiveDirectly-Novissi program (hereafter GD-Novissi), implemented jointly by the Togolese Ministry of Digital Transformation and the NGO GiveDirectly, provided monthly cash transfers to 138,589 individuals in rural parts of Togo between November 2020 and August 2021. Eligible women received 8,620 FCFA (USD $15.50)[3] per month, and eligible men received 7,450 FCFA (USD $13) per month for six months following their registration to the program. GD-Novissi was one of several targeted humanitarian aid programs delivered under the Novissi umbrella during the COVID-19 pandemic.[4]

Like other Novissi programs, registration and payment for GD-Novissi were entirely digital. Individuals registered for the program by dialing a toll-free mobile shortcode and filling out a brief USSD form. Registration required (i) a valid voter ID number, (ii) a valid SIM card, and (iii) access to a mobile phone.[5] Benefits for eligible beneficiaries were delivered monthly via mobile money, with mobile money accounts automatically opened for subscribers who did not already have them.

GD-Novissi used both geographic and poverty-related criteria for eligibility determina-

[3] 8,620 FCFA about = USD$38 (PPP-adjusted).

[4] Other programs included three months of cash transfers of $16-19 delivered to 567,002 informal workers in urban areas between April and August 2020, three months of transfers of $16-19 delivered to 5,850 residents of the canton of Soudou in August 2020, and one-time payments of $8-10 to 244,302 residents of the Savanes region in February 2021. Each of these Novissi programs was linked to lockdowns in the area of eligibility.

[5] A single SIM card could only register one voter ID, but multiple voter IDs could be registered on a single mobile device. Based on household survey data, Aiken et al. (2022b) estimate that 65% of individuals and 85% of households in Togo owned a mobile phone in 2019. Aiken et al. (2022b) estimate that 87% of Togolese adults possessed a voter ID.

tion. First, beneficiaries had to be registered to vote in one of the 100 poorest cantons in the country.[6] Second, poverty estimates for each registered subscriber were derived from their pre-program mobile phone records; only subscribers estimated to be living on less than $1.25/day (the poorest 29% of subscribers) were eligible for the program. Aiken et al. (2022b) provides a full description and evaluation of GD-Novissi's targeting approach, and a more systematic discussion of the extent to which these eligibility criteria created systematic exclusions from the program.

GD-Novissi was advertised over a number of channels, including radio advertisements in eligible geographies, communication with community leaders, and active outreach by small field teams. Mobile subscribers inferred to be living in eligible geographies and below the poverty cut-off were also contacted via SMS messages encouraging them to register for the program. The program launched in November 2021; after three months GD-Novissi had received 181,028 registrations, of which 49,083 met the eligibility criteria, had an active mobile phone, and received benefits.

2.1 RCT Design and Implementation

Our main analysis focuses on a randomized controlled trial (RCT) implemented among all 49,083 individuals who registered for GD-Novissi in November 2020, December 2020, or January 2021 and met the poverty and location-based eligibility criteria. Prior to registration, subscribers who met the eligibility criteria were randomly assigned to treatment (N=27,673) and control (N=21,410) groups. Upon registration, subscribers in the treatment group immediately received the first of their five monthly cash transfers of 8,620 FCFA (USD $15.50) for women or 7,450 FCFA (USD $13) for men. The last of the subscribers in the treatment group, therefore, received their final transfer at the end of May 2021. Subscribers in the control group received the same total benefits as the treatment group, but their payments were distributed beginning in June 2021 and were bundled into three payments (N=21,410). Subscribers in the control group were not informed that they would receive transfers at a later date; the June payments came as a surprise.

[6]The GD-Novissi program's first phase targeted the poorest 100 cantons of Togo and later phases expanded to the poorest 200 cantons; our focus is on the first phase of the program.

2.2 Survey Data

2.2.1 Endline Survey

To evaluate the impact of GD-Novissi cash transfers, we conducted a large phone survey with both treatment and control groups in May 2021, between zero and two months after members of the treatment group had received their final cash transfer (and before any of the control group subscribers had received a transfer — see Figure S1 for a visualization of the project timeline). This 'endline' survey took approximately 30 minutes, and was conducted by enumerators from the Togolese National Institute of Statistics and Economic and Demographic Studies (INSEED). To avoid introducing response biases, surveyors were told not to mention GD-Novissi until the end of the survey and were not aware of the respondent's status (control or treatment). For each call, the enumerator first asked to speak to the person registered for GD-Novissi by name. If that person was unavailable, then the surveyor attempted to complete the survey with the person who answered the phone. In 83% of the cases, the respondent declared being the person who was registered with GD-Novissi.

Sample frame. The sample frame for the endline survey was drawn from subscribers who successfully enrolled in the GD-Novissi RCT between November 2020 and January 2021 (N=49,083). Thus, the sample was restricted to individuals who (a) had active mobile phone accounts; (b) were registered to vote in one of Togo's 100 poorest cantons; (c) completed the registration procedure for GD-Novissi; and (d) were predicted, based on their mobile phone data, to consume less than $1.25 per day. As we discuss later, the homogeneity of this population has important implications for our downstream analysis.

The sample was stratified by treatment status and geography. The former was done to maximize statistical power in estimating treatment effects; the latter was done to account for the fact that one large region (Savanes) received payments unrelated to GD-Novissi during the period when GD-Novissi benefits were being delivered.[7]

Response rate. The enumerators called all 24,294 phone numbers of our final sample in random order. We successfully surveyed 10,129 individuals (response rate of 42%). After removing low-quality surveys (see Appendix A), our final sample contained 9,511 observa-

[7]See Appendix C.2 for details on the other Savanes program. Of the 36,090 subscribers registered in the Savanes region, we sampled 36%; of the 12,993 subscribers registered outside of the Savanes region, we sampled 88%.

tions (completion rate of 39%). This completion rate is similar to other phone surveying completed during COVID-19: for example Egger et al. (2021) analyzes random-digit dialing to conduct surveys on well-being in nine countries during COVID-19 and reports completion rates ranging from 17% to 59%. Table S1 shows that attrition rates do not differ statistically significantly between the treatment and control groups.

Weights. To make our analysis representative of the entire population of eligible beneficiaries, our main results weight observations by the inverse of the probability of being sampled and the inverse of the probability of response. Sampling probability is determined by the four sampling strata (in vs. outside of the Savanes region and treatment vs. control). Response weights are derived from a machine learning model trained to predict response from pre-survey covariates among the 24,294 phone numbers sampled for the survey. The pre-survey covariates include administrative information from the Novissi database and pre-survey phone records. More details on weights in the impact evaluation survey are included in Appendix A.2.

Outcomes. The survey contained modules on household food security and consumption, health, access to social services, poverty, mental health, and experience with the Novissi program. Following our pre-analysis plan (American Economic Association Registry #7590), we constructed seven primary indexed outcomes using the index construction methodology described in Bryan et al. (2021). These seven indices are: food security, financial health, financial inclusion, mental health, perceived socioeconomic status, labor supply, health care access, and labor supply.[8]

We standardize all our outcomes so that, within our control group of eligible active mobile subscribers, each outcome had zero mean and unit variance, with the exception of the mental health measure for which we use the Kessler K6 distress scale methodology (Kessler et al., 2002). For all indexed outcomes variables, we first standardize each component — signed coherently beforehand — by subtracting its control group mean and dividing by its control group standard deviation. We then calculate the sum of the standardized components and standardize the sum again by the control group standard deviation.[9] In addition to these seven primary welfare outcomes, we collected a proxy-means test (PMT), which provides a

[8]Table S3 reports the specific wording of each component of each outcome index.

[9]We impute missing components using the other components in an index unless the missing components are children-related and the family had no children, in which case we compute the index omitting those components.

very rough estimate of consumption, measured in US dollars per day.[10]

Summary Statistics and Balance Checks. Table S2 provides summary statistics and balance checks for the impact evaluation sample. We observe small and generally statistically insignificant differences in treatment assignment by gender, age, occupation, and place of residence, evaluated using the endline survey (Panel A). We separately test for balance using administrative data obtained from the Novissi program (Panel B), and observe no statistically significant imbalance in covariates between treatment and control groups.

2.2.2 Pre-treatment Survey

While much of our analysis — and in particular, the impact evaluation conducted with survey data — relies primarily on the endline survey conducted post-treatment, portions of our analysis use a pre-treatment phone survey conducted in September 2020, prior to the roll-out of the GD-Novissi program. This sample frame was defined as all active mobile subscribers whose primary home location was in those 100 poorest cantons, using geographic information available in the mobile phone data (see Appendix B). In total, we completed 9,484 pre-treatment surveys.

As the primary objective of the pre-treatment survey was to collect PMT data that could be used to train the machine learning algorithms used to identify eligible GD-Novissi beneficiaries (Aiken et al., 2022b), it was different from the endline survey in two key respects. First, it was shorter and more focused on the PMT; it did not contain a mental health module, and there were fewer food security questions (Table S5). Second, the population was designed to be representative of *all* active mobile phone subscribers in Togo's 100 poorest cantons (i.e., the regions that were eligible for GD-Novissi), not just those subscribers predicted to be below the poverty threshold (as in the endline survey). As can be seen in the first two columns of Table S2, the pre-treatment sample was still quite poor (average estimated daily per capita consumption of \$1.49, with a \$0.74 standard deviation), but less homogeneously poor than in the endline survey (average consumption \$1.31, SD \$0.49). Additional details on the pre-treatment survey are provided in Appendix B.

[10]The PMT contains 12 components selected based on joint correlation with consumption in a nationally representative household survey conducted in Togo in 2018. Details of PMT construction are documented in Aiken et al. (2022b); PMT components and weights are listed in Table S4.

2.3 Mobile Phone Metadata

We obtained comprehensive mobile phone metadata from Togo's two mobile network operators in two six-month periods that match to the pre-treatment survey (March - September 2020) and to the endline survey (November 2020 - April 2021). These data include detailed metadata about each phone call and text message sent or received on the mobile networks, including the phone number of the caller and recipient, the timestamp, the duration (for calls only), and the cell tower through which the call was placed. The data also include mobile data usage, including the phone number of the subscriber, the timestamp, and the amount of mobile data used for each mobile data transaction.[11]

We obtained informed consent from each respondent in the pre-treatment and endline surveys to match their survey responses to their mobile phone records.[12] We then generated sets of mobile phone *features* describing how each survey respondent used their mobile phone in the period preceding the survey. Features were generated using open source library `cider`[13] following the procedure described in Aiken et al. (2022b). In total, we constructed 824 features relating to calling patterns, contact networks, mobility, location, data usage, international transactions, and more. Features for training models on the pre-treatment survey were generated using six months of mobile phone data preceding the survey (i.e., April - September 2020); features for training models on the endline survey were generated using mobile phone data from the six-month treatment period (November 2020 - April 2021).

3 Program Impacts Estimated using Survey Data

Our first set of results uses the endline survey to estimate the causal impact of GD-Novissi. These results are based on weighted regressions of each of the seven outcomes on treatment status and include strata, enumerator, and week of the survey fixed effects. To account for multiple hypotheses, we include p-values adjusted for the False Discovery Rate (Anderson, 2008) for our seven pre-specified outcome indices.

Results in Table 1 Panel A indicate that GD-Novissi increases food security (by 0.06 standard deviations, $p = 0.003$), mental health (by 0.07 standard deviations, $p < 0.001$),

[11] Although the dataset shared by the mobile network operators also includes records of mobile money use, we do not use mobile money transactions in our main analysis since the treatment itself was delivered via mobile money and thus mechanically (and dramatically) changed mobile money usage patterns for the treatment group. However, we explore the inclusion of mobile money data in Section 5.1.

[12] Following the data protection procedures described in our IRB protocol, we pseudonymized or removed all personally identifying information, including phone numbers, prior to linking these two datasets.

[13] https://global-policy-lab.github.io/cider-documentation/

and self-perceived socioeconomic status (by 0.04 standard deviations, $p = 0.074$). Consistent with the evidence on the effect of cash transfers in other contexts (Banerjee et al., 2017, 2022), GD-Novissi does not decrease individual labor supply (the coefficient is positive but not statistically significant, with a point estimate close to zero). Although the coefficient estimates are positive, we observe no statistically significant effects on our indices of financial health, financial inclusion, or healthcare access.[14] The last column of Table 1 Panel A indicates that GD-Novissi increases an aggregate welfare index by 0.06 standard deviations ($p = 0.008$), where the aggregate index is constructed as an aggregated normalized index of the seven underlying outcome indices. For reference, the first column of Table 1 shows that there were no statistically significant impact on the proxy means test (PMT) measure of wealth. This is perhaps unsurprising given that the PMT is comprised of components related to long-term poverty, which we do not expect would be influenced by a relatively modest cash transfer (Table S4).

In Appendix C, we test for treatment effect heterogeneity on four pre-registered dimensions: gender, poverty, occupation, and region of residence (in or outside of the Togo's northernmost region, Savanes). To summarize, we find little evidence that treatment effects were heterogeneous across any of these dimensions. The one exception is geography: we find treatment effects on food security and mental health were significantly larger for beneficiaries in the Savanes region in the far North of Togo (Figure S2 and Table S6). Appendix C discusses this geographic heterogeneity in greater detail.

4 Program Impacts Estimated Using Mobile Phone Data

Here we present our main, null result: when using mobile phone data to estimate the treatment effect of GD-Novissi, most estimates are close to zero and statistically insignificant — despite the fact that the survey-based estimates were broadly positive and statistically significant. An advantage of mobile phone data is that, in principle, they could help predict outcomes for a very large population (i.e., the full population of beneficiaries and non-beneficiaries with phones), using only a small sample survey to train the prediction model.

[14] Our financial inclusion index measures the fraction of bank accounts and mobile money usage in households, excluding mobile money accounts of the respondents. GD-Novissi automatically opened mobile money accounts for beneficiaries without an account. According to our survey, 91% of the participants in the control group hold a mobile money account. By providing a new mobile money account in the household, GD-Novissi reduces the return to opening another new account by someone else, which can help explain the absence of effects on this outcome.

If successful, such an approach could enable new paradigms for more rapid and lower cost impact evaluation using digital data. However, we find that in this context, mobile phone data do not capture the same treatment effects estimated with survey data.

We test the potential for using mobile phone metadata for impact evaluation under two different regimes. In the first regime, we assume that the only opportunity for survey data collection occurs before treatment is administered. In this scenario, pre-treatment survey data (collected in September 2020) are matched to pre-treatment phone data (March - September 2020), and machine learning is used to predict socioeconomic status from phone data – this is the machine learning procedure that the GD-Novissi program used to determine program eligibility, as described by Aiken et al. (2022b). Then, after the program has been implemented, an impact evaluation is conducted by comparing the *predicted* post-program outcomes of treated and control individuals, where the predicted outcomes are generated by applying the trained model (trained on pre-treatment data) to phone data collected post-program.[15]

In the second regime, we instead assume that the only opportunity to collect survey data is after treatment has been administered. In this approach, the machine learning algorithm is trained on post-treatment data (i.e., endline survey data from May 2021 that are matched to post-treatment phone data from November 2020 - April 2021), for a small sample of the actual beneficiary population. The trained model is then used to predict outcomes for all subscribers enrolled in the RCT, including those not surveyed.

4.1 Predicting Welfare Levels from Phone Data

We begin by testing the extent to which machine learning models can accurately predict individual welfare, separately in the first regime of analysis (using pre-treatment surveys) and the second regime (using endline surveys). These results are summarized in Table 1 Panels B and C.

Panel B of Table 1 reports the accuracy with which pre-treatment survey outcomes can be predicted from mobile phone data. Specifically, for each of the outcome indices captured in the pre-treatment survey, we calculate the five-fold cross-validated R^2 as follows. The full dataset that matches completed surveys to phone records (N=8,899) is divided randomly into five partitions ("folds"). A machine learning model is trained on four of the five folds and predictions are produced for the observations in the remaining fold; the process is repeated

[15] This regime is similar in spirit to that used in the impact evaluations based on satellite imagery explored in Huang et al. (2021) and Ratledge et al. (2021).

for each of the remaining four folds. The percentage of variation explained by the predictions (R^2) is then calculated, pooling predictions across all the folds.[16] We use a gradient boosting machine learning model, with hyperparameters tuned using nested cross-validation for each of the five folds.[17]

Predictive accuracy is highest for the proxy means test (PMT) poverty index ($R^2 = 0.143$, or a pearson correlation coeffecient of $r = 0.381$) – a finding that replicates results previously documented in Aiken et al. (2022b). However, this performance is worse than the $r = 0.46$ obtained when Aiken et al. (2022b) trained their model using a nationally representative in-person household survey. It is also lower than the $r = 0.68$ found by Blumenstock et al. (2015), when using nationally representative phone surveys from Rwanda. Most critically, we find that we cannot accurately predict any of the other welfare indices (food security, financial health, perceived status, and labor supply) from mobile phone features, in the analysis regime based on the pre-treatment survey ($R^2 = 0.002$ - 0.046).[18]

Panel C of Table 1 shows results from analogous experiments, where the machine learning model is instead trained using the endline survey data. With this survey, the predictive accuracy for the PMT is much lower ($R^2 = 0.049$, or a pearson correlation coefficient of 0.253) than it was with the pre-treatment survey. This difference is likely due in part to the more homogeneous population represented in the endline survey: while both surveys focus on the same rural regions of the country, the endline survey was also restricted to *program-eligible* mobile subscribers, where one criteria for eligibility was that they were predicted to be poor (see Section 2.2.1). As with the pre-treatment survey, we observe little predictive accuracy for food security, mental health, perceived socioeconomic status, or any of the other welfare indices ($R^2 = 0.002$ - 0.026).

To summarize: we find that in these relatively homogeneous samples, ML models perform very poorly at predicting a wide range of welfare outcomes from mobile phone metadata. The main exception is with the PMT, where the models explain 5-14% of the variation in poverty. Importantly, however, the machine learning model cannot accurately predict any welfare index for which statistically significant treatment effects were observed in the endline survey (in Table 1). We discuss the implication of these limitations in Section 5 after

[16] Survey weights and response weights are used in both training and calculating R^2 scores.

[17] Hyperparameters are: (1) Winsorization of features (selected from {no winsorization, 1% winsorization}, (2) minimum data in each leaf of the forest (selected from {10, 20, 50}), (3) number of leaves for each tree (selected from {5, 10, 20}), (4) learning rate (selected from {0.05, 0.075, 0.1}), and (5) number of trees (selected from {50, 100, 200}).

[18] We cannot include results on financial inclusion, mental health, healthcare access, nor the seven-index composite from Table 1, as the questions required to construct these indices were not included in the pre-treatment survey.

presenting our primary analysis of treatment effects.

4.2 Estimating Treatment Effects from Phone Data

Our next set of tests investigates whether the phone-based estimates of welfare described in Section 4.1 can be used to directly estimate the treatment effects of GD-Novissi. In the first regime of analysis – where the only survey data available are collected pre-treatment – we train the machine learning model on the pre-treatment survey (as evaluated in Panel B of Table 1). We then compare *predicted endline outcomes* of treated and control individuals, where predicted endline outcomes are generated by passing post-treatment mobile phone data through the prediction model that was trained pre-treatment.[19]

Panel D of Table 1 shows the predicted average treatment effect of GD-Novissi, which is obtained by regressing the predicted outcome on the individual's treatment status. To estimate the variance of treatment effects derived from these phone-based predictions of welfare, we use a Bayesian bootstrap procedure (Rubin, 1981) to incorporate both the first-stage uncertainty in our ML models' predictions of each outcome, as well as the variation in predictions across treatment and control subscribers.[20] The treatment effect is not statistically significant for any of the welfare indices besides the PMT, where it is negative. For food security and perceived socioeconomic status — which both had significant treatment effects in the survey — the point estimates of phone-based treatment effects are -0.003 and -0.013, respectively.

Panel E of Table 1 shows the results from the second regime where only endline survey data are available to train the machine learning model. Here, the models are trained using endline data (as evaluated in Panel C of Table 1); the models are then used to generate predicted welfare outcomes for all subscribers using post-treatment mobile phone data.[21] We estimate the predicted average treatment effects as before by comparing predictions between the treatment and control groups, and estimates of variance are again produced

[19] Specifically, one model corresponding to each welfare outcome is trained using the pre-treatment survey and phone data from before treatment (March - September 2020), with hyperparameters tuned through 5-fold cross-validation specific to that outcome. Each model is then used to generate predicted welfare outcomes for treated and control individuals, using mobile phone data collected from the period during which treatment is administered (November 2020 - April 2021).

[20] Following Angrist et al. (2017) and Dobbie and Song (2020), we assign each observation that appears in the training and/or inference sets a "bootstrap weight" drawn from a Dirichlet distribution Dirichlet(1, ..., 1). These weights are used (in combination with the survey and response weights) in training the ML model, and in calculating treatment effects. We repeat this procedure with 100 different random draws, and report the mean and standard deviation across these 100 bootstrap estimates.

[21] These subscribers include the surveyed subscribers which the model is trained on; results are unchanged if we restrict the inference set to subscribers who were not surveyed and therefore not used in training.

with a Bayesian bootstrap. We do not estimate statistically significant treatment effects on food security, financial inclusion, perceived socioeconomic status, or the combined index (all of which showed statistically significant impacts in the survey-based impact evaluation); phone-based point estimates for these treatment effects range from 0.001 to 0.015 standard deviations. There is a positive and statistically significant treatment effect estimated for the mental health index (0.028 standard deviations, $p = 0.053$) and healthcare access index (0.021 standard deviations, $p = 0.076$).

4.3 Additional Tests of Robustness

In Section 5, we posit and examine several reasons why the positive treatment effects estimated using survey data were, in general, not observed in predictions generated from mobile phone data. First, however, we present a few tests to ensure that the preceding results are robust to different variations of the machine learning methodology used to generate predicted treatment effects.

First, we find that results are unchanged if we vary the period during which mobile phone data are used to generate predictions of endline welfare. This experiment addresses the possibility that mobile phone data may be most impacted by cash transfers in the period immediately following the transfer. For this test, we train and evaluate prediction models that use only two weeks of mobile phone data (instead of the six months used in our main analysis). When matching to the pre-treatment survey, we use mobile phone data from the two weeks during which the survey was conducted (September 17-30, 2020); for the post-treatment period, we use data from the two weeks immediately following the date on which each individual registered for GD-Novissi (which is also the date the first payment was delivered). In Table S7 we do not observe improved predictive performance relative to Table 1 (R^2 = -0.002 - 0.112), and estimated treatment effects are similar in magnitude and significance.

Second, we test whether using *changes* in mobile phone use between the pre-treatment period and the post-treatment period can improve predictive performance. Changes are calculated, for each feature, as the simple difference between a subscriber's value in the post-treatment period and the value in the pre-treatment period; if the feature in either period is missing then the difference is also missing. We then train a model using the endline survey as ground truth to predict each of our outcomes from these measures of change. As before, we produce predictions for each subscriber enrolled in the RCT but not surveyed, and compare predictions for the treatment and control groups. Table S8 shows that using

changes in phone use as features does not improve predictive performance for our machine learning models, for either a six-month or two-week period for deriving the pre-treatment and during-treatment features (R^2 = -0.004-0.036).

Third, we test whether impact estimates based on mobile phone data are significant on subsets of the population where the survey-based treatment effects were largest. In particular, as discussed in detail in Appendix C, the survey data indicate that treatment effects on food security and mental health are larger for beneficiaries in the Savanes region in the north of Togo. However, Table S9 shows that when the machine learning model is trained using data from survey respondents in Savanes, and then evaluated on treatment and control individuals only in Savanes, the corresponding treatment effects estimated from the phone data remain statistically insignificant. Note that we do not find improved predictive power by restricting to just mobile subscribers in Savanes (R^2 = -0.012 - 0.120). Thus, it is not surprising that the treatment effects derived from phone data in the Savanes region are likewise not significant.

Finally, we run several tests of the machine learning models themselves. In addition to tuning the hyperparameters as described above, we take additional steps to ensure that data are not too sparse for the specific models being used (Bellman and Kalaba, 1959). Specifically, we introduce a feature selection step prior to model fitting, which eliminates all phone-based features that are not statistically significantly different between the treatment and control groups. Using a two-tailed t-test, we find that 180 features (22% of all features) are different between the treatment and control groups at a 0.05 level in the six-month period. Despite the substantial share of features that differ systematically between treatment and control subscribers, this feature selection step does not improve predictive accuracy (Table S10; R^2 = 0.000 - 0.139) and does not impact the statistical significance of treatment effects.

5 Discussion

To summarize our main results, we find that (i) using endline survey data, we estimate positive and statistically significant impacts of GiveDirectly-Novissi cash transfers on food security, financial inclusion, mental health, perceived socioeconomic status, and an aggregate outcomes index; (ii) using mobile phone data, we can accurately predict a PMT-based measure of poverty in the cross-section, but we cannot accurately predict the outcomes that were impacted by the program; (iii) as a likely result, the treatment effects of GD-Novissi estimated using mobile phone data are not statistically significant.

The first result is broadly consistent with a number of studies demonstrating positive impacts of cash transfers on food security and mental health during the COVID-19 pandemic (Banerjee et al., 2020; Bottan et al., 2021; Londoño-Vélez and Querubin, 2022; Karlan et al., 2022). In comparison to other papers on COVID-19 cash transfers, the GD-Novissi transfer size is slightly smaller (monthly transfers USD 13-15.5 in comparison to USD 15-52 in other studies). However, effect sizes are of a similar magnitude to those observed in other studies (in the range of 0.04-0.07 standard deviations).[22]

The subsequent results are more nuanced, and inform a rapidly evolving debate about if and how new digital data sources can be used to inform development research and policy. Where several recent studies have shown that phone data and machine learning can produce accurate estimates of consumption and asset-based wealth, we find that — at least in the rural Togolese context — a similar procedure does not produce reliable estimates of food security, mental health, or self-perceived economic status.

5.1 Challenges to estimating vulnerability indices from mobile phone data

Here we explore four hypotheses that could explain why mobile phone data and machine learning can accurately predict wealth, but do not accurately predict food security or the other self-reported welfare outcomes.

5.1.1 Noise in survey data

Survey-based measures of food security and other vulnerability outcomes may be noisier than survey-based measures of economic poverty, and therefore more difficult to estimate from any underlying data source (Hjelm et al., 2016; Tadesse et al., 2020). However, the survey-based indices have low enough measurement error that we are able to identify statistically significant survey-based treatment effects on a number of outcomes from the GD-Novissi cash transfers (Table 1), suggesting that noise in ground truth is not a primary contributor to the low predictive power of our models. To further test this hypothesis, we train classification

[22]Our results on heterogeneous treatment effects are also broadly consistent with other papers studying the impacts of COVID-19 cash transfers on well-being, which for the most part do not find significant heterogeneity across dimensions studied (Londoño-Vélez and Querubin, 2022; Karlan et al., 2022). However, two results stand in contrast: Londoño-Vélez and Querubin (2022) finds that treatment effects are driven primarily by households in urban areas, while we find that treatment effects are driven primarily by households in Savanes, which is the most rural region of Togo; and Karlan et al. (2022) finds that treatment effects on food security are larger for female-headed households than male-headed households, whereas we find no heterogeneous treatment effects by gender of the recipient.

models to predict an outcome with little to no measurement error: *treatment status*. Since we managed the treatment assignment ourselves (see Aiken et al. (2022b)), we know which phone numbers received cash transfers and which didn't. The results in Table S11, which are produced using the same machine learning pipeline (implemented with gradient boosting trees as described in Section 4.1), indicate that mobile phone data does not accurately predict GD-Novissi treatment status either in the country as a whole (AUC = 0.515 - 0.522) or in the Savanes region only where we find larger treatment effects in the survey data (AUC = 0.516 - 0.518). This result suggests measurement error in the survey is not the main reason for the null results in Table 1.[23]

5.1.2 Population homogeneity

A second possible explanation for the low predictive power of the food security models — in comparison to past work on predicting in large cross-sections of populations of mobile phone subscribers — is the homogeneity of the study sample. Past work on predicting poverty in Togo and elsewhere has typically focused on identifying variation in poverty across an entire country, or across large regions (Aiken et al., 2022b; Blumenstock et al., 2015; Blumenstock, 2018). In comparison, we seek to identify variation in poverty and vulnerability measures within a homogeneous subset of the population, focusing on either the population of the poorest 100 cantons of Togo (when conducting analysis with the pre-treatment survey) or a sub-population identified to be living in poverty within those poorest 100 cantons (when conducting analysis with the endline survey). In past work that has compared the accuracy of predicting poverty from mobile phone data in full-country evaluations vs. in rural areas only, predictive power is typically substantially lower when restricting to rural areas only (r = 0.46 vs. 0.31 in Togo for poverty prediction at the individual level (Aiken et al., 2022b) and r = 0.64 vs. 0.50 in Rwanda for poverty prediction at the district level (Blumenstock et al., 2015)).[24]

[23]To further confirm this result — and to test the validity of our pipelines for machine learning with mobile phone data — we replicate the experiment of predicting treatment status from six months of mobile phone data during the treatment period, this time including 'cheat code' features relating to mobile money use in the machine learning model. These features include information on the number and sizes of transactions placed and received by each subscriber, and thus directly reveal information about whether a subscriber has received a GD-Novissi cash transfer via mobile money. With the mobile money-related features included, the area under the curve score for predicting treatment status is 0.998 (evaluated over five-fold cross-validation). This helps confirm the integrity of the machine learning methodology. Table S12, which shows the feature importances for this machine learning model, further confirms that the key features used by the model relate to mobile money transactions.

[24]Related work has shown that satellite imagery has higher predictive accuracy for village-level poverty in full-country evaluations than in evaluations restricted to rural areas (Jean et al., 2016; Chi et al., 2022).

However, while the homogeneity of the population explains why predictive power for the proxy-means test is lower than in previous papers that evaluate nationally representative samples, it does not explain why predictive power for all vulnerability indices is substantially lower than for the PMT in both the pre-treatment and endline survey.

5.1.3 Relationship between phone use and vulnerability

A third hypothesis for why we are unable to predict any of our vulnerability indices from mobile phone data (when we are, to some extent, able to predict poverty) is that the mobile phone use may be more closely related to long-term poverty outcomes than to short-term vulnerability metrics. For example, mobile money and mobile data usage are important predictors of wealth in Aiken et al. (2022b), and are related to long-term investments in smartphones and financial services technologies. Short-term changes in food security, financial health, and mental health may not result in the types of investments in phone capabilities (such as buying a new smartphone or investing in a large airtime bundle) that would be clearly observable from mobile phone metadata. On the other hand, prior work has shown that phone use changes significantly in response to short-term shocks (Bagrow et al., 2011; Blumenstock et al., 2016); we might therefore expect that other signals in the phone data, such as the timing or volume of outgoing calls, would reflect short-term changes in welfare. We cannot directly differentiate between these hypotheses in our context, but believe it is an interesting area for future work.

5.1.4 Spatial structure in outcome indices

A fourth and final hypothesis is that long-term poverty may have more geographic structure than food security or the other vulnerability-related outcomes we examine. Spatial features obtained from the locations of cell towers through which subscribers place calls are critical features in the phone-based poverty prediction models described in Aiken et al. (2022b). It is possible that food security and other vulnerability indices are less predictable from mobile phone data because they are less related to geographic information. To test for whether spatial structure could explain the difference in predictive power for the PMT in comparison to vulnerability indices, in Table 2, we calculate the within and between variance grouped by canton for both the pre-treatment and endline survey. We find that the ratio of

For instance, Yeh et al. (2020) report an R^2 of 0.70 for full-country analysis and 0.32 for within-country analysis, averaged across several African nations. Aiken et al. (2023) likewise find that, across 10 countries, satellite-based poverty predictions consistently have a substantially higher rank correlation with true poverty at a national scale than in urban and rural areas separately.

between variance to within variance is substantially higher for the PMT (8.2 - 15.0 in the pre-treatment and endline surveys) than for any of the vulnerability indices (1.3 - 2.1 across surveys and indices). This result, combined with past documentation that spatial structure plays a key role in estimating poverty from mobile phone data (Aiken et al., 2022b; Hernandez et al., 2017), suggests that spatial structure in an index may be a key determinant of whether it can be predicted from mobile phone data.

5.2 Additional challenges to estimating treatment effects from mobile phone data

Even if it were possible to accurately predict welfare outcomes from phone data in the cross-section, it might still prove difficult to use phone data to estimate the treatment effects of cash transfers on those same outcomes. Here, we provide suggestive evidence of two such issues: that the modest size of the GD-Novissi cash transfers may generate only small changes in phone use, and that model drift in the relationship between phone use and vulnerability may complicate the repeated use of machine learning models over time.

5.2.1 Magnitude of impacts

A challenge for detecting treatment effects from mobile phone data in the context of the Novissi program is the modest impact on the outcomes of interest resulting from relatively small transfer sizes. Our survey-based impact evaluation results detect treatment effects of 0.04-0.07 standard deviations for the outcomes of interest (food security, financial inclusion, mental health, perceived socioeconomic status, and the aggregate welfare index) resulting from six monthly transfers of USD 13-15. In comparison, interventions of a larger magnitude would be expected to produce larger impacts: for example, Haushofer and Shapiro (2016) reports a 0.26 standard deviation increase in both food security and psychological well-being in the year following a large (USD 404-1,525 PPP) cash transfer, and Banerjee et al. (2021) reports a 0.18 standard deviation increase in food security and a 0.12 standard deviation improvement in mental health a year and a half after a "big push" intervention (including a large livestock transfer, training, savings access, and 30 weekly cash transfers of USD 7.6 PPP).

The modest cash transfer sizes and impacts of the GD-Novissi program is reflected in similarly modest impacts on mobile phone use. Table 3 shows the impact of cash transfers on 10 easily interpretable metrics of mobile phone use (selected from among the 823 features used

in the prediction model). The impacts — which range from 0.01 to 0.06 standard deviations in absolute value — are very small relative to the cross-sectional variation in these aspects of phone use. Thus, while we observe statistically significant impacts on 20-32% of features describing mobile phone use, the impacts are small in magnitude. If treatment does not substantially affect phone use, then it would naturally be challenging to study treatment effects from patterns of phone use.

5.2.2 Model drift

A specific challenge to identifying treatment effects in the first regime we study — that is, by training a model prior to program roll-out and deploying it later on to monitor program impacts — is *model drift* in the relationship between features of mobile phone use and ground truth measures of vulnerability over time. Particularly in the context of large economic shocks like the COVID-19 pandemic, a model trained well before a program's implementation may no longer be accurate when cash transfers are distributed. Aiken et al. (2022b) empirically study model drift in Togo, and find a substantial drop in accuracy when a model is trained two years prior to its deployment (Spearman correlation of 0.42 at the time of training vs. 0.35 at the time of deployment). To test the extent to which the same issues of model drift are present in this work, we evaluate the accuracy of predictions from our poverty prediction model trained on the pre-treatment survey for generating predictions using mobile phone data from the treatment period. In comparison to the R^2 score of 0.049 for the model trained on the endline survey, the predictions from the model trained on the pre-treatment survey achieve an R^2 of only 0.030, providing suggestive evidence of model drift in the nine months that elapsed between the pre-treatment and endline surveys.

6 Conclusion

Our results show that in a context where survey data indicate small but statistically significant cash transfer impacts on several welfare outcomes, impact estimates derived from mobile phone records are not statistically significant. These results suggest that machine learning predictions of welfare derived from digital sources – while effective for estimating regional poverty (Blumenstock et al., 2015) and targeting policies (Aiken et al., 2022b) – cannot naively replace traditional survey-based measurements in program monitoring and impact evaluation. Mobile phone data may perform better when treatment effects are larger, when there is more heterogeneity in key outcomes in the beneficiary population, and when

phone-based estimates of outcomes are more accurate. More generally, our results suggest that estimating *changes* in wealth using mobile phone data and other digital traces is more challenging than estimating *levels* of wealth.

References

Aiken, E., Bedoya, G., Blumenstock, J., and Coville, A. (2022a). Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in afghanistan. *arXiv preprint arXiv:2206.11400*.

Aiken, E., Bellue, S., Karlan, D., Udry, C., and Blumenstock, J. E. (2022b). Machine learning and phone data can improve targeting of humanitarian aid. *Nature*, 603(7903):864–870.

Aiken, E., Rolf, E., and Blumenstock, J. (2023). Fairness and representation in satellite-based poverty maps: Evidence of urban-rural disparities and their impacts on downstream policy. *arXiv preprint arXiv:2305.01783*.

Aker, J. C., Boumnijel, R., McClelland, A., and Tierney, N. (2011). Zap it to me: The short-term impacts of a mobile cash transfer program. *Center for Global Development working paper*, (268).

Anderson, M. L. (2008). Multiple inference and gender differences in the effects of early intervention: A reevaluation of the abecedarian, perry preschool, and early training projects. *Journal of the American Statistical Association*, 103(484):1481–1495.

Angrist, J. D., Hull, P. D., Pathak, P. A., and Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics*, 132(2):871–919.

Bagrow, J. P., Wang, D., and Barabási, A.-L. (2011). Collective Response of Human Populations to Large-Scale Emergencies. *PLoS ONE*, 6(3):e17680.

Banerjee, A., Duflo, E., and Sharma, G. (2021). Long-term effects of the targeting the ultra poor program. *American Economic Review: Insights*, 3(4):471–86.

Banerjee, A., Faye, M., Krueger, A., Niehaus, P., and Suri, T. (2020). Effects of a universal basic income during the pandemic. *Innovations for Poverty Action Working Paper*.

Banerjee, A., Karlan, D., Trachtman, H., and Udry, C. R. (2022). Does poverty change labor supply? evidence from multiple income effects and 115,579 bags. *National Bureau of Economic Research*, 27314.

Banerjee, A. V., Hanna, R., Kreindler, G. E., and Olken, B. A. (2017). Debunking the stereotype of the lazy welfare recipient: Evidence from cash transfer programs. *The World Bank Research Observer*, 32(2):155–184.

Barriga-Cabanillas, O., Blumenstock, J. E., Lybbert, T., and Putman, D. (n.d.). The potential and limitations of big data in development economics: The use of cell phone data for the targeting and impact evaluation of a cash transfer program in haiti? *Presentation at 2021 Pacific Development Conference*.

Bastagli, F., Hagen-Zanker, J., Harman, L., Barca, V., Sturge, G., Schmidt, T., and Pellerano, L. (2016). Cash transfers: what does the evidence say. *A rigorous review of programme impact and the role of design and implementation features. London: ODI*, 1(7).

Bellman, R. and Kalaba, R. (1959). A mathematical theory of adaptive control processes. *Proceedings of the National Academy of Sciences*, 45(8):1288–1290.

Blumenstock, J., Cadamuro, G., and On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, 350(6264):1073–1076.

Blumenstock, J. E. (2014). Calling for Better Measurement: Estimating an Individual's Wealth and Well-Being from Mobile Phone Transaction Records. In *The 20th ACM Conference on Knowledge Discovery and Mining (KDD 2014), Workshop on Data Science for Social Good*, New York, NY.

Blumenstock, J. E. (2018). Estimating economic characteristics with phone data. In *AEA papers and proceedings*, volume 108, pages 72–76.

Blumenstock, J. E., Eagle, N., and Fafchamps, M. (2016). Airtime Transfers and Mobile Communications: Evidence in the Aftermath of Natural Disasters. *Journal of Development Economics*, 120:157–181.

Bottan, N., Hoffmann, B., and Vera-Cossio, D. A. (2021). Stepping up during a crisis: The unintended effects of a noncontributory pension program during the covid-19 pandemic. *Journal of Development Economics*, 150:102635.

Bryan, G., Choi, J. J., and Karlan, D. (2021). Randomizing religion: the impact of protestant evangelism on economic outcomes. *The Quarterly Journal of Economics*, 136(1):293–380.

Chetty, R., Friedman, J. N., and Rockoff, J. E. (2014). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review*, 104(9):2593–2632.

Chi, G., Fang, H., Chatterjee, S., and Blumenstock, J. E. (2022). Microestimates of wealth for all low-and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3):e2113658119.

Dobbie, W. and Song, J. (2020). Targeted debt relief and the origins of financial distress: Experimental evidence from distressed credit card borrowers. *American Economic Review*, 110(4):984–1018.

Doocy, S. and Tappis, H. (2017). Cash-based approaches in humanitarian emergencies: a systematic review. *Campbell Systematic Reviews*, 13(1):1–200.

Egger, D., Miguel, E., Warren, S. S., Shenoy, A., Collins, E., Karlan, D., Parkerson, D., Mobarak, A. M., Fink, G., Udry, C., et al. (2021). Falling living standards during the covid-19 crisis: Quantitative evidence from nine developing countries. *Science Advances*, 7(6):eabe0997.

Fatehkia, M., Tingzon, I., Orden, A., Sy, S., Sekara, V., Garcia-Herranz, M., and Weber, I. (2020). Mapping socioeconomic indicators using social media advertising data. *EPJ Data Science*, 9(1):22.

Gilraine, M., Gu, J., and McMillan, R. (2020). A new method for estimating teacher value-added. Technical report, National Bureau of Economic Research.

Haushofer, J. and Shapiro, J. (2016). The short-term impact of unconditional cash transfers to the poor: experimental evidence from kenya. *The Quarterly Journal of Economics*, 131(4):1973–2042.

Hernandez, M., Hong, L., Frias-Martinez, V., Whitby, A., and Frias-Martinez, E. (2017). Estimating poverty using cell phone data: evidence from guatemala. *World Bank Policy Research Working Paper*, (7969).

Hidrobo, M., Hoddinott, J., Peterman, A., Margolies, A., and Moreira, V. (2014). Cash, food, or vouchers? evidence from a randomized experiment in northern ecuador. *Journal of development Economics*, 107:144–156.

Hjelm, L., Mathiassen, A., and Wadhwa, A. (2016). Measuring poverty for food security analysis: consumption-versus asset-based approaches. *Food and nutrition bulletin*, 37(3):275–289.

Huang, L. Y., Hsiang, S. M., and Gonzalez-Navarro, M. (2021). Using satellite imagery and deep learning to evaluate the impact of anti-poverty programs. Technical report, National Bureau of Economic Research.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., and Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794.

Kane, T. J. and Staiger, D. O. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. Technical report, National Bureau of Economic Research.

Karlan, D., Lowe, M., Osei, R. D., Osei-Akoto, I., Roth, B. N., and Udry, C. R. (2022). Social protection and social distancing during the pandemic: Mobile money transfers in ghana. *National Bureau of Economic Research working paper*.

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S.-L., Walters, E. E., and Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological medicine*, 32(6):959–976.

Londoño-Vélez, J. and Querubin, P. (2022). The impact of emergency cash assistance in a pandemic: experimental evidence from colombia. *Review of Economics and Statistics*, 104(1):157–165.

Lopez, J. (2021). The case of the solidarity income in colombia: The experimentation with data on social policy during the pandemic. *COVID-19 FROM THE MARGINS*, page 126.

Mukerjee, A. N., Bermeo, L. X., Okamura, Y., Muhindo, J. V., and Bance, P. G. A. (2023). Digital-first Approach to Emergency Cash Transfers: Step-kin in the Democratic Republic of Congo. *World Bank Working Paper Series*, (181798).

Ratledge, N., Cadamuro, G., De la Cuesta, B., Stigler, M., and Burke, M. (2021). Using satellite imagery and machine learning to estimate the livelihood impact of electricity access. Technical report, National Bureau of Economic Research.

Rubin, D. B. (1981). The bayesian bootstrap. *The annals of statistics*, pages 130–134.

Schwab, B. (2013). In the form of bread? a randomized comparison of cash and food transfers in yemen.

Sheehan, E., Meng, C., Tan, M., Uzkent, B., Jean, N., Burke, M., Lobell, D., and Ermon, S. (2019). Predicting economic development using geolocated wikipedia articles. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2698–2706.

Smythe, I. S. and Blumenstock, J. E. (2022). Geographic microtargeting of social assistance with high-resolution poverty maps. *Proceedings of the National Academy of Sciences*, 119(32):e2120025119.

Sundsøy, P., Bjelland, J., Reme, B.-A., Iqbal, A. M., and Jahani, E. (2016). Deep learning applied to mobile phone data for individual income classification. In *2016 International Conference on Artificial Intelligence: Technologies and Applications*, pages 96–99. Atlantis Press.

Tadesse, G., Abate, G. T., and Zewdie, T. (2020). Biases in self-reported food insecurity measurement: A list experiment approach. *Food Policy*, 92:101862.

Warren, R., Aiken, E., and Blumenstock, J. (2022). Note: Home location detection from mobile phone data: Evidence from togo. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS)*, pages 685–692.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., and Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in africa. *Nature communications*, 11(1):1–11.

Table 1: Survey-based and phone-based treatment effects of the GD-Novissi program

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Survey-based treatment effects									
Treatment	0.002	0.064***	0.026	0.007	0.072***	0.040*	0.010	0.009	0.061***
	(0.012)	(0.022)	(0.024)	(0.021)	(0.019)	(0.022)	(0.023)	(0.025)	(0.023)
Obs.	8,452	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511
Control mean	1.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
FDR q-value		0.014	0.428	0.732	0.002	0.149	0.732	0.732	0.021
Panel B: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.143	0.002	0.013	—	—	0.034	—	0.046	—
Obs.	8,899	8,899	8,899	—	—	8,899	—	8,890	—
Panel C: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.049	0.008	0.009	0.003	0.002	0.008	0.007	0.021	0.026
Obs.	8,448	9,507	9,507	9,134	9,507	9,507	9,522	9,507	9,507
Panel D: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.007***	-0.003	-0.013	—	—	-0.013	—	-0.000	—
	(0.002)	(0.016)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	48,759	48,759	48,759	—	—	48,759	—	48,759	—
Control Mean	1.409	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.731	0.016	0.159	—	—	0.483	—	0.002	—
Panel E: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.005	0.001	0.004	0.028*	0.001	0.021*	0.011	0.015
	(0.002)	(0.013)	(0.018)	(0.020)	(0.017)	(0.017)	(0.015)	(0.013)	(0.015)
Obs.	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759
Control Mean	1.314	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.800	0.021	0.404	0.910	0.088	0.162	0.677	0.955	0.100

Notes: Panel A shows treatment effects of GD-Novissi estimated using the endline survey. The dependent variable for each regression is indicated in the column title; see Appendix A for variable construction. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling and response probabilities. Panels B and C show the performance of the machine learning models used to generate the predictions of welfare outcomes used to generate phone-based estimates of the treatment effects, measured by R^2 score (evaluated out-of-sample over five fold cross validation on the training set).Panels D and E report the treatment effects of GD-Novissi derived using the phone-based machine learning model's predictions. In Panel D, the pre-treatment survey is used to train the machine learning model; in Panel E, the endline survey is used to train the model. In Panels D and E treatment effects are estimated across all subscribers enrolled in the RCT by regressing the phone-based estimate of the outcome variable on treatment status, with standard errors determined with a Bayesian bootstrap procedure. The Z-test p-values in Panels D and E indicate the significance of the Z-test that the phone-based treatment effect and the survey-based treatment effect reported in Panel A are different. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table 2: Variation Between and Within Canton

Outcome	(1) Between Variance	(2) Within Variance	(3) Ratio
Panel A: Pre-treatment survey			
PMT	4.30	0.29	15.00
Food Security	1.438	0.96	1.44
Financial Health	1.25	0.98	1.27
Perceived Socioeconomic Status	1.80	0.95	1.89
Labor Supply	1.58	0.94	1.67
Panel B: Endline survey			
PMT	2.08	0.25	8.21
Food Security	1.85	0.99	1.87
Financial Health	1.65	0.99	1.87
Financial Inclusion	1.59	0.92	1.72
Mental Health	1.89	0.98	1.92
Perceived Socioeconomic Status	1.26	0.99	1.28
Healthcare Access	1.46	1.00	1.45
Labor Supply	1.57	0.96	1.63
Aggregate Welfare Index	2.03	0.98	2.08

Notes: Between vs. within variance, with groups defined by canton (self-reported in the baseline survey, determined by location of Novissi registration in the endline survey). Only individuals in cantons with at least 10 individuals surveyed are included in the analysis. All outcomes except for the PMT are standardized to 0 mean and unit variance in the control group.

Table 3: Treatment Effects on Phone Use

	(1) Active days	(2) Calls	(3) Texts	(4) Contacts	(5) International Contacts	(6) % Initiated	(7) Regions	(8) Prefectures
Treatment	0.064***	0.021**	0.010	0.033***	0.015	-0.026***	0.049***	0.038***
	(0.009)	(0.009)	(0.010)	(0.009)	0.015 (0.013)	(0.010)	(0.009)	(0.009)
Control Mean	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Unstandardized Control Mean	98.892	625.654	70.979	77.098	3.164	0.902	4.236	9.867
Obs.	48,803	48,664	44,548	48664	22,410	44,548	48,803	48,803

Notes: Treatment effects on basic metrics of mobile phone use, selected from among the 823 metrics of mobile phone use used by our machine learning models. Metrics were selected by hand from the pool based on ease of interpretation. All features are standardized to zero mean and unit variance in the control group (unstandardized control mean is also provided for intuition). $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

A Additional Details of the Impact Evaluation Survey

A.1 Data Collection Monitoring

We identified surveyors who performed poorly by comparing the data collected with the information contained in Novissi administrative data. We began our analysis by constructing "enumerator effects" (EE) estimates for every enumerator in our data. We predicted the EE on the basis of the correct answers to five questions for which we obtained the "truth" from the Novissi registry (prefecture, canton, age, gender, and Novissi status), and on the frequency of very short surveys (below 15 minutes) as well as surveys with no children reported (which avoids the roster part of the survey and simplifies the surveyor's work.). We controlled for interviewee characteristics such as region and interview language to separate the enumerator's impact from observable interviewee selection.[25] Our approach to estimating EE parallels the parametric empirical Bayes estimator of teacher effects (Kane and Staiger (2008); Chetty et al. (2014); Gilraine et al. (2020)).

We then normalized the EEs for each of the seven dimensions (prefecture, canton, age, gender, Novissi status, number of short surveys, number of surveys without kids), and took the sum of the coherently signed components for enumerators who conducted more than ten interviews. We classified the interviews of enumerators with an average EE lower than the sample mean minus two standard deviations as of "very poor quality" and remove them from the sample. 615 observations collected by five enumerators who were ranked "very poor quality" were removed from the dataset. In addition, on the second day of the survey, while monitoring data quality, we noticed an enumerator who was performing extremely poorly. After a warning from his supervisor, the quality of his data collection improved. We removed the data collected by this enumerator during the first two days (60 observations). Thus, we only use data from only 9,511 high-quality surveys in our main analysis.

A.2 Weights

We reweight observations in the endline survey by the inverse of the sampling probability and the inverse of the probability of response. Sampling probabilities are determined by the four sampling strata, as follows:

- Savanes, treatment: 30.48%

[25] The phone number list was randomized and then distributed to the enumerators, so we believed that there is little room for sorting.

- Savanes, control: 41.26%
- Outside Savanes, treatment: 76.25%
- Outside Savanes, control: 100.00%

To calculate response weights, we train a machine learning model to predict survey response from pre-survey covariates. In total, 9,511 phone numbers completed the survey out of 24,294 numbers sampled. We include the following pre-survey covariates as features in our model:

- 824 features relating to phone use in the six months pre-survey (November 2022 - April 2022).
- 6 features from the Novissi registry: Age, gender, canton of registration (one hot encoded), number of payments received up until the survey date, profession (one hot encoded for the 20 most common professions), and registration week (one hot encoded).
- An indicator for treatment vs. control group.

Using a similar pipeline to the machine learning methods described in Section 4.1, we train a LightGBM classifier to predict response, and produce an out-of-sample predicted probability of response for each phone number using five-fold cross-validation. We tune hyperparameters using three-fold cross-validation separately on each of the five folds. With all predictions pooled together, our model achieves an AUC of 0.69. In Figure S3, we confirm that the predicted probabilities of response are well-calibrated by binning the predicted probability of response into ten equal-sized bins, and plotting the average realized response rate for phone numbers in each bin.

The overall weight for each observation is the product of the inverse of the sampling probability and the inverse of the (predicted) probability of response.

A.3 Outcome Construction

We constructed our seven primary outcomes using the index construction methodology from Bryan et al. (2021). Specifically, we first standardized each component by subtracting its control group mean and dividing by the control group standard deviation. We then calculated the sum of the standardized components, and standardized the index again by subtracting the control group mean and dividing by the control group standard deviation. As specified in our

pre-analysis plan, our seven primary indices were formed with the components documented in Table S3.

A.4 PMT

We collected information to calculate a proxy-means test (PMT) for each subscriber that proxies for consumption. We used the proxy-means test developed in Aiken et al. (2022b), which used machine learning methods to select twelve features that are most jointly predictive of consumption (training on data from a nationally representative household survey conducted in Togo in 2018). The weights for the resulting linear model are listed in Table S4.

A.5 Attrition and Balance Checks

We test for differential nonresponse between the treatment group and the control group in the impact evaluation survey by regressing a binary indicator for response on treatment status, among all 24,294 phone numbers called. In Table S1, we find that there is no statistically significant difference in response rates between the treatment and control groups.

We also test for covariate balance between the treatment and control groups in our impact evaluation survey sample in Table S2. We find that the treatment and control groups in the impact evaluation survey are balanced on self-reported age, gender, and occupation (Panel A), with results robust to substituting administrative data from the Novissi program for self-reported survey data (Panel B).

B Additional Details of the Pre-Treatment Survey

Details of the sampling and design for the pre-treatment survey — conducted pre-program in September 2020 (see Figure S1) — are available in Aiken et al. (2022b).

The phone survey reached 9,484 mobile subscribers inferred to be living in parts of Togo eligible for the GD-Novissi program, and collected information on poverty, living conditions, food security, and health. Critically, the pre-treatment survey collected the components for our PMT (Appendix A) and the components of three of the four indices for which we observe significant treatment effects of GD-Novissi: food security, financial health and perceived socioeconomic status.

The financial health and perceived socioeconomic status indices are constructed identically to the indices in the impact evaluation survey; however, only certain components of the food security index were collected in the pre-treatment survey, so we construct a "reduced food security index" for the pre-treatment survey. The reduced food security index is less comprehensive than the food security index collected in the impact evaluation survey (Table S3) and, in particular, does not include questions on food consumption. The components for the reduced food security index are listed in Table S5.

As in the impact evaluation survey, each index is constructed following Bryan et al. (2021), by standardizing each component across the surveyed population, summing components, and then standardizing the resulting index.

Also, as in the impact evaluation survey, each observation is weighted by the inverse of the sampling probability and the inverse of the probability of response. Details on the estimation of weights are available in Aiken et al. (2022b). We use weights throughout our analysis involving the pre-treatment survey, except where otherwise noted.

C Treatment Effect Heterogeneity

We test for treatment effect heterogeneity on four pre-registered dimensions: gender, poverty, occupation, and region of residence (in or outside of the Togo's northernmost region, Savanes). For each dimension, we test for heterogeneous treatment effects on our seven outcomes and the aggregate welfare index in Table 1 Panel A.

C.1 Heterogeneity by gender, wealth, and occupation

We find little evidence that treatment effects were heterogeneous across the socioeconomic and demographic subgroups that we pre-specified in our pre-analysis plan. In particular, while GD-Novissi had an important gender component, whereby women received roughly 15% more money per month than men, the welfare impacts on women were not significantly larger than for men. These results can be seen in Panel A of Table S6: while women, in general, are worse off than men (the third row of coefficients), the coefficient on the interaction between treatment and female is never significantly different from zero.[26]

Panels B and C of Table S6 likewise indicate that treatment effects did not differ by pre-treatment wealth or occupation. In Panel B, we compare treatment effects for people with PMT scores above and below the sample median, and, with the exception of healthcare access, do not observe significant differences for any outcome. Panel C indicates that treatment effects were not different for farmers – the most common occupation in the rural areas of Togo where GD-Novissi was implemented, and the occupation reported by 60% of endline survey respondents.

C.2 Geographic heterogeneity

There is, however, one dimension where we find evidence of substantial heterogeneity in treatment effects, which is by the location of the beneficiary. In particular, Panel D of Table S6 highlights how treatment effects on food security and mental health were significantly larger for beneficiaries in the Savanes region in the far North of Togo (Figure S2). Indeed, with the exception of healthcare access, the treatment effects for beneficiaries outside Savanes are all close to zero and no longer statistically significant once we account for the differential effect of treatment in Savanes.

[26] According to administrative data from the Novissi program, women represent half of GD-Novissi beneficiaries, and 45% of our surveyed sample. However, in suggestive evidence of strategic behavior at the household level in GD-Novissi registration, the share of women among survey respondents is only 27%. That is, 40% of the phone numbers registered with female voter ID cards were answered by men.

Savanes is unique in several respects: most GD-Novissi beneficiaries (70%) reside in the Savanes region, it is generally poorer than other regions eligible for GD-Novissi, it had higher rates of COVID-19 and related curfews than other regions, and the government provided an independent round of cash transfers called *Savanes-Novissi* (unconnected to GD-Novissi) to all residents in Savanes in February 2021 (two months before our endline surveys were conducted). While understanding the reasons for the substantial geographic heterogeneity between Savanes and non-Savanes beneficiaries is not the focus of this paper, we explore briefly below three possible hypotheses that could explain why the treatment effects of GD-Novissi are observed mainly in the Savanes region: (i) that differential registration for Savanes-Novissi between the GD-Novissi treatment and control groups resulted in additional cash impacts for the treatment group, (ii) that mobility reductions resulting from curfews in the Savanes region made cash transfers more impactful in Savanes than the rest of the country, and (iii) that price differences between Savanes and the rest of the country gave cash transfers more purchasing power in Savanes.

C.2.1 Interaction Between GD-Novissi and Savanes-Novissi

In addition to the GD-Novissi program considered here, the Government of Togo implemented three other targeted cash transfer programs under the Novissi umbrella during the pandemic period. One of these, "Savanes-Novissi", provided one-time cash transfers of USD 8-10 to all residents of Savanes who registered for Novissi in a two-week period beginning on February 22, 2021. Women received a one-time transfer of CFA 6,125 (USD 9.80), and men received a transfer of CFA 5,250 (USD 8.40). A total of 244,302 Savanes residents registered for and received Savanes-Novissi, of whom 114,311 (46.79%) were already registered for GD-Novissi.

We observe an approximately 20 percentage point difference in registration rates for Savanes-Novissi between the treatment and control groups in GD-Novissi, with the control group substantially more likely to register for the Savanes-Novissi program. 41% of the treatment group registered for Savanes-Novissi, while 63% of the control group registered.

There are two plausible explanations for the difference in enrollment: first, GD-Novissi provided enough assistance for the treatment group, so they were less in need of further cash transfers, and second, that confusion in communications around the two programs resulted in members of the treatment group believing they were ineligible for Savanes-Novissi. The second explanation is particularly plausible for two reasons. First, people located in Savanes who were registered for GD-Novissi were eligible for Savanes-Novissi, but were required to

register separately for Savanes-Novissi, which could be confusing. Second, because the treatment group was receiving cash transfers from GD-Novissi in February 2021, the government of Togo initially excluded treated people of GD-Novissi from the Savanes-Novissi program. While the Savanes-Novissi amount was transferred at the time of registration for everyone else, people in the GD-Novissi treatment group received Savanes-Novissi cash transfers in the second week of the period of registration.

We included specific questions in the impact evaluation survey to distinguish between the two explanations. We first asked people if they registered with Savanes-Novissi. If not, we asked them an open-ended question why not (Table S13). The enumerators were told to classify the answers in one of the eight pre-defined categories, including "other" and "I don't know". Three of the possible categories are related to the confusion hypothesis ("I did not think I was eligible", "I did not think I needed to register, since I already registered with GD-Novissi", and "I heard about the program after the end of the registration period"). Three others are related to GD-Novissi impact hypothesis ("I receive GD-Novissi, I don't need extra money", "I have enough money, I don't need extra money", and "Other people are more in need than me, I prefer them to get the money").

Qualitatively, the treatment group is more likely to be confused about the eligibility criteria than the control group. The first main reason why people did not register is the lack of information, and there is a ten percentage points difference between the treatment and the control group: 36.5% of the control group versus 49.6% of the treatment group was confused about the eligibility criteria. Less than 3% of people in both treatment arms reported a lack of need for Savanes-Novissi as the main reason, supporting the second hypothesis for the enrollment differences (confusion in eligibility criteria).

However, in comparing the welfare outcomes of people who did and did not register with Savanes-Novissi by treatment arm (Table S14), we do observe that people from the treatment group who did not register with Savanes-Novissi have a higher food security and financial health index than those who did register. There are no such differences between registrants and non-registrants in the control group. The fact that the treatment group self-selected in Savanes-Novissi supports the first hypothesis, suggesting that GD-Novissi contributed to the low enrollment rates for Savanes-Novissi in the treatment group.

We conclude, based on this evidence, that both hypotheses (the welfare impact of GD-Novissi and confusion around eligibility criteria) likely contributed to lower registration rates for GD-Novissi in the treatment group in Savanes. Importantly, however, the difference in registration rates does not explain the larger welfare impacts of GD-Novissi in Savanes in

comparison to the rest of the country: if anything, we would expect the lower registration rates for Savanes-Novissi in the treatment group to attenuate treatment effects relative to other regions of the country.

C.2.2 Impacts of the Savanes Curfew on Mobility

After a surge in the number of COVID-19 cases, the entire region of Savanes was placed under curfew from January 17 to February 21, 2021. It is the only part of Togo where there was a strict shutdown due to COVID-19 during the implementation of GD-Novissi. A plausible explanation for the welfare impacts of GD-Novissi in Savanes is that mobility restrictions led to an increased need for assistance in the Savanes region relative to other parts of the country.

We test this hypothesis using mobility indicators derived from mobile phone data from subscribers inferred to be living in the Savanes region before and during the period of curfew. Using the frequency-based home location detection methods described in Warren et al. (2022) we infer which subscribers are likely to be residing in the Savanes period during the curfew. We proxy mobility with the number of unique towers and the number of unique cantons visited over the course of 21 days in November, December, January, and February.[27]

First, we use the phone-based mobility indicators to verify whether the curfew in the Savanes region induced people to move less. We use a difference-in-differences strategy to identify the effect of the curfew on our sample mobility. Our estimating OLS equation is

$$y_{it} = \rho_i + \sum_{\tau=-3, \tau \neq -1}^{0} \beta_\tau (S_i \times C_\tau) + \gamma_t + \alpha X_{it} + \varepsilon_{it}$$

where ρ_i is an individual fixed-effect that captures all observable and unobservable time-invariant individual characteristics, γ_t is a period fixed-effect, and ε_{it} is an individual-period shock. S_i is a dummy for living in Savanes, and C_τ $(-3 \leq \tau \leq 0)$ are dummies indicating the number of periods relative to the curfew (February 2021). We control for the number of transactions X_{it} to make sure that the potential change in our mobility metric is not mechanically driven by a change in the number of phone calls and texts. The parameter of interest is β_0, which measures the effect of the curfew on our metric for mobility y_{it} relative to the previous period (the omitted category β_{-1}). The coefficients β_{-3} and β_{-2} are pre-

[27] Whenever someone makes a call or text, that transaction is associated with the tower closest to where she is. However, many people in our sample do not make a transaction every day. The number of unique towers or cantons observed in the data is the best information we have on actual mobility, but remains a noisy metric for mobility.

trends coefficients that capture the difference in mobility between Savanes and Not-savanes inhabitants relative to the omitted variable before the curfew.

Table S15 shows the curfew's effect on mobility. Mobility in the Savanes region decreases by 0.3 towers per month on average (compared to a control mean of 5.76 towers per month), and the pre-trend coefficients are not statistically different from zero, indicating that the curfew put in place in the Savanes region did significantly decrease mobility of the region's residents.

To test whether mobility reductions may have driven the positive GD-Novissi treatment effects in Savanes, we test for differential impacts of GD-Novissi among Savanes subscribers by baseline mobility quintiles (with baseline mobility derived pre-treatment in the months of September and October 2021). Mobility is again derived from mobile phone data and proxied by the number of unique cell towers a subscriber visits in the months of September and October. While we observe substantial mobility reductions across quintiles (Figure S4), we do not observe a differential impact of GD-Novissi by baseline mobility (Figure S5). This result suggests that it is unlikely that mobility impacts drove the GD-Novissi treatment effects in Savanes, as there is no differential impact for the most mobile pre-curfew within the Savanes region.

C.2.3 Price Differences

A final possible explanation for the GD-Novissi treatment effects in Savanes (in comparison to the rest of the country) is that price differences between the Savanes region and the rest of the country give GD-Novissi transfers more purchasing power in Savanes. We collected price information for staple goods in the consumption module of the impact evaluation survey; in our analysis, we restrict to goods for which at least 50% of the respondents provided a price. Among these seven goods, we observe statistically significant differences in prices between Savanes and the rest of the country for only three goods (Table S16). Palm oil and milk are more expensive in Savanes, while Niebe is cheaper. Given that there are no systematic price differences in a consistent direction between Savanes and the rest of the country, we conclude that price differences are not a major driver of the GD-Novissi treatment effects in Savanes.

D Additional Tests for Estimating Treatment Effects from Mobile Phone Data

In this section, we use alternative specifications to test whether it is possible to recover treatment effects from GD-Novissi mobile phone records. In table S7 we test using a two-week period to derive features from mobile phone data rather than a six-month feature period. In Table S8, we try using changes in features between the pre-treatment and during-treatment periods to predict each of our outcomes (using the endline survey as ground truth). In Table S9 we try predicting outcomes and inferring treatment effects in the Savanes region only, since the survey-based treatment effects were only observable in Savanes. In Table S10 we try the same specifications using only features that are statistically significantly different between the treatment and control groups (22% of all features). Finally, to test for whether noise in survey data is the cause of low predictive power, in Table S11 we train and evaluate a model to predict treatment status from the mobile phone feature set. The poor performance of each of these models suggests that it is the inability of phone data to identify differences between the treatment and control groups — rather than an issue of noisy survey data — that drives the low predictive power of the phone-based models and thus the null effects in downstream inference tasks.

Supplementary Figures and Tables

Table S1: Differential attrition

	Probability of non-response
Treatment	-0.01 (0.01)
N	24,294
Control Mean	0.61

Notes: Differential attrition from impact evaluation survey sample for treatment and control groups. Effect of treatment on attrition is estimated with a simple linear regression specification with non-response as the dependent variable. Regression is conducted without fixed effects. $^{*}p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table S2: Summary Statistics and Balance Checks

	Baseline Sample		*Endline Sample*		
	N	Mean	N	Mean	Diff. T-C
Panel A. Survey data					
PMT	8,821	\$1.49	8,452	\$1.31	\$0.00
		(0.74)		(0.49)	(0.01)
Female	8,821	0.23	9,511	0.31	0.03**
		(0.42)		(0.46)	(0.01)
Age	8,716	33.37	9,310	36.03	-0.30
		(11.98)		(11.44)	(0.30)
Farmers	8,819	0.41	9,511	0.59	-0.02
		(0.49)		(0.49)	(0.01)
Savanes	8,821	0.51	9,443	0.72	-0.01
		(0.50)		(0.45)	(0.01)
Panel B. Novissi registry data					
Female	5,493	0.50	9,511	0.49	0.02*
		(0.50)		(0.50)	(0.01)
Age	5,493	36.02	9,429	37.63	0.09
		(13.96)		(12.70)	(0.33)
Farmers	5,402	0.23	9,375	0.38	-0.02*
		(0.42)		(0.49)	(0.01)
Savanes	5,493	0.52	9,511	0.74	-0.00
		(0.50)		(0.44)	(0.01)

Notes: This Table presents summary statistics for the pre-treatment and the endline samples. Column "N" indicates the number of respondents, "Mean" indicates the sample mean, with the standard deviation in parenthesis. Column "Diff. T-C" contains the balance checks that are conducted by regressing the demographic variable of interest on treatment status (balance checks are conducted for the endline survey only). All observations are weighted by sampling probabilities. All regressions control for the enumerator, week of the survey, and strata fixed effects. Robust standard errors are in parenthesis. $^{*}p$ <0.1; $^{**}p$ <0.05; $^{***}p$ <0.01.

Table S3: Components for impact evaluation outcomes

Question	**Possible Answers**
Panel A. Food security	
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of a lack of money or other resources?	Two weeks
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often did you have to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children over 3 in your household had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children over 3 in your household eat?	0-3
When was the last time your household had each of the following items: Powdered milk, sugar, smoked anchovy, fresh onion, dried fish, sesame, red palm oil, traditional bread, orange, cowpea/dried beans.	0 Never - 5 Less than a week
How much did you spend on purchasing each of the items above, the last time?	Integer
Panel B. Financial health	
Were you able to save money last month? If so, how much?	Integer
God forbid, if your household stopped getting income from any source, how long could your household easily continue to meet your basic needs for food and housing? (Winsorized 95th percentile.)	Integer
God forbid, if there was a major emergency and your household needed money, how much money could you easily obtain within the next seven days? (Winsorized 95th percentile.)	Integer
Panel C. Financial inclusion	
Fraction of the adults in the household with a bank account.	Float
Fraction of the adults in the household (excluding the respondent) with a mobile money account.	Float
Panel D. Mental health (Kessler K6 nonspecific distress scale)	

During the past 7 days, about how often did you feel nervous?	Integer
During the past 7 days, about how often did you feel hopeless?	Integer
During the past 7 days, about how often did you feel restless or fidgety?	Integer
During the past 7 days, about how often did you feel that everything was an effort?	Integer
During the past 7 days, about how often did you feel so sad that nothing could cheer you up?	Integer
During the past 7 days, about how often did you feel worthless?	Integer
Panel E. Self perception of socioeconomic status	
In general, relative to other people in Togo, would you say that you are...	1 very poor - 5 very well off
How do you think other communities perceive the wealth of your household?	1 very poor - 5 very well off
Panel F. Labor supply	
Hours worked last week (winsorized 99th percentile)	Integer
During the past 7 days, how much income/pay did you receive?	Integer
Panel G. Healthcare access	
The last time you or someone else in your household needed healthcare, did you get healthcare?	Yes/no
When you last needed health care, did you get it at the hospital?	Yes/no
God forbid, if a child in your household needed to go the hospital, would you be able to bring him or her?	Yes/no

Notes: Components for each of the outcomes in the endline survey. All indices are produced using the index construction methadology from Bryan et al. (2021) except for the mental health index, which is based on simple addition of the components.

Table S4: Proxy-means test

Component	Weight	Component	Weight	Component	Weight	Component	Weight
Car	2.77	Pref. Tchaudjo	0.09	HHW educ. 4	-0.18	Pref. Amou	-0.34
Stove	1.77	Pref. Bassar	0.07	Pref. Lacs	-0.18	HHW educ. 3	-0.34
Refridgerator	1.32	Pref. Haho	0.04	Pref. Sotouboua	-0.18	Pref. Plaine du Mo	-0.34
HHH educ. 8	1.12	Pref. Dankpen	-0.03	Pref. Kloto	-0.21	Pref. Anie	-0.34
HHH educ. 9	0.91	Pref. Moyen-Mono	-0.06	HHW educ. 6	-0.21	Pref. Tandjoare	-0.35
Hospitalization	0.81	Pref. Oti-Sud	-0.08	Pref. Kpele	-0.21	Pref. Binah	-0.37
Iron	0.63	Pref. Oti	-0.11	Pref. Bas-Mono	-0.23	Pref. Ave	-0.39
HHH educ. 3	0.55	Pref. Wawa	-0.11	Pref. Lome	-0.23	Pref. Keran	-0.41
TV	0.50	Pref. Vo	-0.12	Pref. Danyi	-0.24	Pref. Kpendjal	-0.46
Children in school	0.48	Pref. Ogou	-0.14	Pref. Yoto	-0.26	HHW educ. 2	-0.50
Pref. Cinkasse	0.39	Pref. Tone	-0.15	Pref. Agoe-Nyive	-0.27	Pref. Kozah	-0.51
Pref. Tchamba	0.33	Pref. Agou	-0.17	HHH educ. 5	-0.27	HHH educ. 2	-0.57
Toilet	0.26	Pref. Akebou	-0.17	No children in school	-0.31	Pref. Blitta	-0.61
HHH educ. 7	0.17	HHW educ. 1	-0.17	Pref. Assoli	-0.32	HHH educ. 1	-0.63
Pref. Est-Mono	0.14	Number of children	-0.17	Pref Kpendjal-Ouest	-0.33	Pref. Golfe	-0.68
HHW educ. 0	0.12	HHW educ. 4	-0.18	Pref. Zio	-0.33	Pref. Doufelgou	-0.75

Notes: Weights for proxy-means test (PMT) determined using Ridge regression. See Aiken et al. (2022b) for details of PMT feature selection and construction.

Table S5: Reduced food security index

Question	Possible Answers
Yesterday, how many meals did you eat?	0-3
In the past 7 days, how often were you unable to eat preferred foods because of lack of money or other resources?	0 Never - 4 Every day
In the past 7 days, how often have you had to limit portion size at meal times?	0 Never - 4 Every day
In the past 7 days, how often have you had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
In the past 7 days, how often have the children in your household over age three had to reduce the number of meals eaten in a day?	0 Never - 4 Every day
Yesterday, how many meals did the children in your household over age three eat?	0-3
In the past 7 days, were you able to buy the amount of food you usually buy?	Yes/no

Notes: Components for the reduced food security index in the pre-treatment survey.

Table S6: Survey-based treatment effect heterogeneity

	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Health care access	(7) Labor supply	(8) All seven indices
Panel A: Gender								
Treatment * Female	-0.037	-0.015	-0.070	0.006	0.003	-0.075	-0.011	-0.053
	(0.049)	(0.051)	(0.047)	(0.041)	(0.048)	(0.054)	(0.050)	(0.049)
Treatment	0.077***	0.034	0.025	0.072***	0.042	0.036	0.018	0.081***
	(0.025)	(0.029)	(0.025)	(0.023)	(0.027)	(0.026)	(0.033)	(0.027)
Female	-0.050	-0.121***	0.201***	-0.074**	-0.116***	-0.079**	-0.221***	-0.122***
	(0.035)	(0.039)	(0.034)	(0.030)	(0.037)	(0.037)	(0.037)	(0.036)
Panel B: Poverty								
Treatment * Poor	0.039	-0.021	-0.068	-0.049	-0.021	0.098**	-0.049	-0.019
	(0.045)	(0.052)	(0.046)	(0.039)	(0.047)	(0.048)	(0.054)	(0.048)
Treatment	0.053	0.048	0.040	0.105***	0.051	-0.048	0.024	0.072**
	(0.033)	(0.036)	(0.034)	(0.028)	(0.034)	(0.035)	(0.034)	(0.035)
Poor	-0.120***	-0.024	-0.111***	0.012	-0.080**	0.062*	0.058	-0.054**
	(0.030)	(0.037)	(0.033)	(0.029)	(0.034)	(0.033)	(0.040)	(0.035)
Panel C: Occupation								
Treatment * Farmer	0.016	0.032	0.024	-0.032	-0.027	0.049	0.028	0.024
	(0.046)	(0.050)	(0.044)	(0.039)	(0.046)	(0.048)	(0.052)	(0.048)
Treatment	0.052	0.006	-0.011	0.090***	0.053	-0.018	-0.010	0.043
	(0.037)	(0.039)	(0.035)	(0.031)	(0.036)	(0.039)	(0.043)	(0.039)
Farmers	-0.152***	-0.114***	-0.277***	-0.031	-0.172***	-0.002	-0.135***	-0.234***
	(0.032)	(0.037)	(0.032)	(0.028)	(0.035)	(0.033)	(0.040)	(0.035)
Panel D: Region								
Treatment * Savanes	0.087**	0.041	0.055	0.064*	0.065	-0.076*	0.015	0.066
	(0.042)	(0.044)	(0.041)	(0.037)	(0.044)	(0.045)	(0.043)	(0.043)
Treatment	0.000	-0.003	-0.033	0.025	-0.007	0.066*	-0.002	0.012
	(0.032)	(0.031)	(0.031)	(0.029)	(0.034)	(0.034)	(0.029)	(0.032)
Savanes	-0.076**	0.024	-0.037	0.014	-0.010	0.144***	0.017	0.020
	(0.032)	(0.033)	(0.030)	(0.029)	(0.035)	(0.032)	(0.033)	(0.033)
Obs	9,511	9,511	9,511	9,511	9,511	9,511	9,511	9,511

Notes: Heterogeneous treatment effects for outcomes for which we detect a statistically significant survey-based treatment effect in Table 1 Panel A. The dependent variable for each regression is indicated in the column title; see Appendix A for variable construction. In Panels A, C, and D, gender, occupation, and region of residence are determined by information provided by the respondents in the survey. In Panel B, poverty is determined by having a below-median PMT score. All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities and response probabilities, and observations are restricted to subscribers who were active prior to the program's launch. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S7: Estimating Treatment Effects from Two Weeks of Mobile Phone Data

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.112	0.003	0.014	—	—	0.017	—	0.028	—
Obs.	8,593	8,593	8,593	—	—	8,593	—	8,584	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.031	0.004	0.006	-0.002	0.004	0.001	0.007	0.010	0.011
Obs.	8,238	9,261	9,261	8,898	9,261	9,261	9,276	9,261	9,261
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.004***	0.001	0.014	—	—	-0.003	—	0.007	—
	(0.001)	(0.014)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	46,327	46,327	46,327	—	—	46,327	—	46,327	—
Control Mean	1.467	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.822	0.018	0.673	—	—	0.745	—	0.005	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.014	0.008	-0.012	-0.001	0.012	0.004	0.011	0.010
	(0.001)	(0.012)	(0.013)	(0.015)	(0.014)	(0.012)	(0.012)	(0.012)	(0.012)
Obs.	46,327	46,327	46,327	46,327	46,327	46,327	46,327	46,327	46,327
Control Mean	1.306	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.793	0.049	0.524	0.464	0.002	0.267	0.819	0.943	0.049

Notes: Replication of Table 1 using two weeks of phone data to derive features (rather than six months). In the first regime — in which the ML models are trained using data from the impact evaluation survey — the two weeks of phone data for model training are obtained from the two weeks during which the pre-treatment survey took place in September 2021. The mobile phone data used to train the ML model in second regime — in which the ML models are trained using data from the endline survey — is taken from the two weeks immediately after each subscriber registered for GD-Novissi. The immediate post-treatment two weeks are used to generate well-being predictions in both regimes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S8: Estimating Treatment Effects from Mobile Phone Data using Changes in Features

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes									
R^2	0.036	0.005	0.004	0.004	-0.004	0.006	0.000	0.015	0.020
Obs.	8,446	9,504	9,504	9,131	9,504	9,504	9,519	9,504	9,504
Panel B: Phone-based treatment effects									
Treatment	-0.003**	0.000	-0.015	-0.007	0.025*	-0.011	0.005	-0.004	-0.005
	(0.002)	(0.018)	(0.019)	(0.016)	(0.017)	(0.018)	(0.019)	(0.015)	(0.014)
Obs.	48,726	48,726	48,726	48,726	48,726	48,726	48,726	48,726	48,726
Control Mean	1.318	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.699	0.025	0.177	0.591	0.062	0.068	0.861	0.657	0.015

Notes: Replication of Table 1 Panel B using changes in phone-derived features between the pre-treatment and during-treatment periods as inputs to the model (once for the six month time period, and once for the two week time period). Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S9: Estimating Treatment Effects from Mobile Phone Data in Savanes

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.120	-0.012	0.009	—	—	0.042	—	0.038	—
Obs.	3,701	3,701	3,701	—	—	3,701	—	3,698	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.034	0.007	-0.003	-0.006	0.001	-0.002	0.001	0.016	0.023
Obs.	3,089	3,478	3,478	3,368	3,478	3,478	3,481	3,478	3,478
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.005***	-0.002	-0.009	—	—	-0.011	—	-0.005	—
	(0.002)	(0.020)	(0.014)	—	—	(0.015)	—	(0.016)	—
Obs.	35,889	35,889	35,889	—	—	35,889	—	35,889	—
Control Mean	1.412	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.778	0.028	0.208	—	—	0.548	—	0.002	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.003**	0.003	0.004	0.010	0.022	-0.008	0.000	0.013	0.008
	(0.002)	(0.018)	(0.018)	(0.024)	(0.020)	(0.022)	(0.017)	(0.017)	(0.016)
Obs.	35,889	35,889	35,889	35,889	35,889	35,889	35,889	35,889	35,889
Control Mean	1.307	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.652	0.034	0.477	0.929	0.068	0.127	0.738	0.900	0.059

Notes: Replication of Table 1 with only subscribers located in Savanes. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S10: Estimating Treatment Effects from Mobile Phone Data using Only Significant Features

	(1) PMT	(2) Food security	(3) Financial health	(4) Financial inclusion	(5) Mental health	(6) Perceived status	(7) Healthcare access	(8) Labor supply	(9) All seven indices
Panel A: Predicting welfare outcomes using ML trained on pre-treatment survey									
R^2	0.139	0.000	0.007	—	—	0.027	—	0.044	—
Obs.	8,899	8,899	8,899	—	—	8,899	—	8,890	—
Panel B: Predicting welfare outcomes using ML trained on endline survey									
R^2	0.042	0.008	0.005	0.002	0.004	0.006	0.008	0.021	0.024
Obs.	8,448	9,507	9,507	9,134	9,507	9,507	9,522	9,507	9,507
Panel C: Phone-based treatment effects trained on the pre-treatment survey									
Treatment	-0.006***	-0.010	-0.013	—	—	-0.013	—	0.001	—
	(0.002)	(0.018)	(0.014)	—	—	(0.013)	—	(0.012)	—
Obs.	48,759	48,759	48,759	—	—	48,759	—	48,759	—
Control Mean	1.431	0.000	0.000	—	—	0.000	—	0.000	—
Z-test p-value	0.755	0.011	0.159	—	—	0.491	—	0.002	—
Panel D: Phone-based treatment effects trained on the endline survey									
Treatment	-0.001	0.001	0.002	0.004	0.028*	-0.001	0.021	0.012	0.016
	(0.002)	(0.015)	(0.018)	(0.019)	(0.017)	(0.017)	(0.016)	(0.013)	(0.015)
Obs.	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759	48,759
Control Mean	1.313	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Z-test p-value	0.802	0.016	0.425	0.920	0.088	0.141	0.709	0.927	0.101

Notes: Replication of Table 1 with only features that are statistically significantly different between the treatment and control groups. Standard errors from Bayesian bootstrap procedure in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S11: Estimating treatment status from mobile phone data

Phone Data Period	**AUC**	*N*
Panel A: All subscribers in RCT		
Six months	0.5151	49,079
Two weeks	0.5219	46,370
Panel B: Savanes only		
Six months	0.5254	36,086
Two weeks	0.5153	34,049

Notes: Predictive performance of a gradient boosting model for predicting treatment status from mobile phone records from the treatment period. Predictions are obtained over 5 fold cross validation, and the pooled area under the curve (AUC) score is reported.

Table S12: Feature importances in a machine learning model including mobile money data

Feature	**Importance**
Maximum amount of transactions in category "other"	281
Maximum balance before outgoing transactions	246
Mean balance before outgoing transactions	211
Maximum balance before outgoing transactions in category "other"	172
Mean amount of transactions in category "other"	145
Number of outgoing transactions	140
Number of outgoing transactions in category "other"	123
Mean balance before outgoing transactions in category "other"	114
Mean balance after outoging transactions in category "other"	105
Maximum balance before outgoing transactions in category "other"	102

Notes: Feature importances for machine learning model predicting treatment status from mobile phone data *including data on mobile money transactions* using six months of phone data from during the treatment period (see Section 5.1). Feature importances are derived from the gradient boosting model as the total number of times a feature is split upon in the entire ensemble of regression trees. Only the top 10 most important features are shown.

Table S13: Reasons for non-registration to Savanes-Novissi

	Control	Treatment
(1) I did not think I was eligible.	16.9% [46]	23.8% [115]
(2) I did not think I needed to register since I registered with GD-Novissi.	5.9% [16]	9.7% [55]
(3) I heard about the program after the end of the registration period.	13.7% [37]	16.1% [80]
(4) I receive GD-Novissi, I don't need extra money.	0.6% [2]	2.2% [13]
(5) I have enough money, I don't need extra money.	1.2% [3]	0% [0]
(6) Other people are more in need than me, I prefer them to get the money.	0.9% [2]	0.2% [1]
(7) Other.	27.8% [68]	16.8% [87]
(8) I don't know or refuse.	33% [78]	30.8% [156]
Total	100% [252]	100% [507]

Notes: Answers to the impact evaluation survey question "Why didn't you attempt to register for Savanes-Novissi in March of this year?", separately between the treatment and control groups (restricted to subscribers who earlier answered that they had not attempted to register for Savanes-Novissi). Observations are weighted by sampling probabilities. Counts are shown in square brackets.

Table S14: Registration with Savanes-Novissi

	(1) Food security	(2) Financial health	(3) Financial inclusion	(4) Mental health	(5) Perceived status	(6) Health care access	(7) Labor supply	(8) **All seven indices**
T, non-SN	0.148***	0.063	0.041	0.091***	0.110***	0.023	0.017	0.131***
	(0.037)	(0.042)	(0.034)	(0.030)	(0.037)	(0.038)	(0.044)	(0.039)
T, SN	0.039	-0.036	0.053	0.070*	0.074*	-0.010	-0.049	0.037
	(0.040)	(0.045)	(0.039)	(0.036)	(0.040)	(0.045)	(0.051)	(0.044)
C, non-SN	0.042	-0.023	0.038	-0.017	0.085**	0.052	-0.036	0.038
	(0.038)	(0.044)	(0.039)	(0.033)	(0.040)	(0.038)	(0.046)	(0.041)
C, SN Mean	-0.01	0.02	-0.01	0.04	0.00	0.04	0.01	0.03
F-test 1-2	7.43***	5.55**	0.10	0.39	0.85	0.57	2.39	5.37**
Obs.	4,755	4,755	4,755	4,755	4,755	4,755	4,755	4,755

Notes: Results for regressing the main survey outcomes on the interaction of GD-Novissi treatment status and Savanes-Novissi registration status. *T* indicates treatment, *C* indicates control, and *SN* and *non-SN* indicates beneficiaries and non-beneficiaries Savanes-Novissi, respectively. *F-test 1-2* row provides the p-value of the statistical comparison of the coefficients for "Treatment, not Savanes-Novissi" and "Treatment, Savanes-Novissi". All regressions control for the enumerator, week of the survey, and strata fixed effects. All observations are weighted by sampling probabilities. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$.

Table S15: Effects of the Savanes curfew on mobility

	Number of unique towers
β_0	-0.305*** (0.120)
β_{-2}	-0.037 (0.117)
β_{-3}	-0.010 (0.119)
N	18,398
Sample Mean	5.76

Notes: Differences-in-differences specification for identifying the impacts of the Savanes curfew on mobility proxied from mobile phone records. β_0 is the coefficient of interest, representing the effect of February (the month of the curfew) relative to January (pre-curfew). β_2 and β_3 represent the effects of November and December (relative to the month of January). *p<0.1; **p<0.05; ***p<0.01.

Table S16: Prices across regions

	(1) Milk (sachet)	(2) Sugar (sachet)	(3) Onion (pile)	(4) Palm oil (liter)	(5) Bread (unit)	(6) Orange (pile)	(7) Niebe (bowl)
Savanes region	50.3*	2.1	-7.7	68.4***	5.3	6.7	-74.9***
	(28.1)	(4.4)	(5.6)	(19.9)	(6.7)	(9.6)	(19.9)
Not-Savanes Mean	352.4	105.3	197.2	875.2	229.5	199.9	1461.5
Obs	1,097	3,200	4,822	2,120	2,524	1,473	2,316

Notes: This table provides price differences between Savanes and non-Savanes regions, based on prices of goods reported by impact evaluation survey respondents. The dependent variables are indicated in the column title. All regressions control for enumerator, week of survey, and treatment status fixed effects. All observations are weighted by sampling probabilities. Robust standard errors are in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Figure S1: GD-Novissi timeline

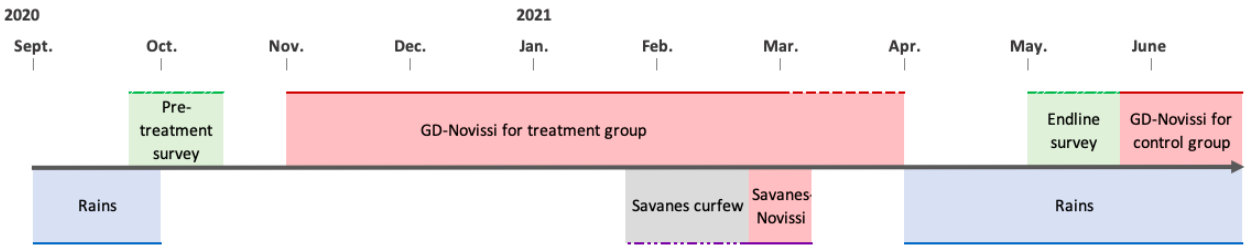

Figure S2: Beneficiaries per canton (admin-3 units)

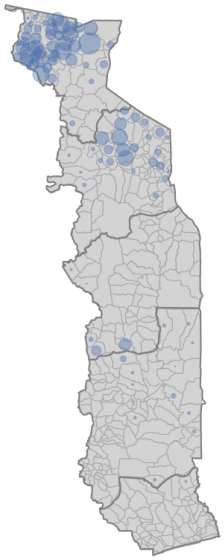

Figure S3: Confirming the calibration of response weights

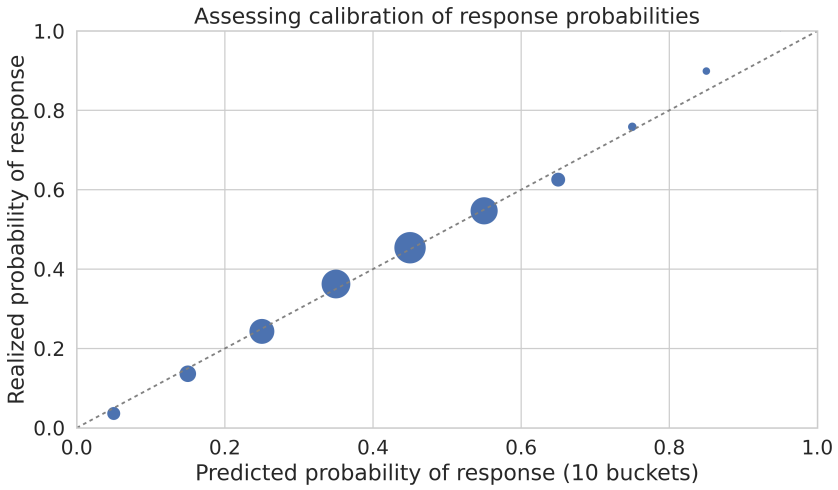

Figure S4: Mobility reductions in Savanes by baseline mobility quintile

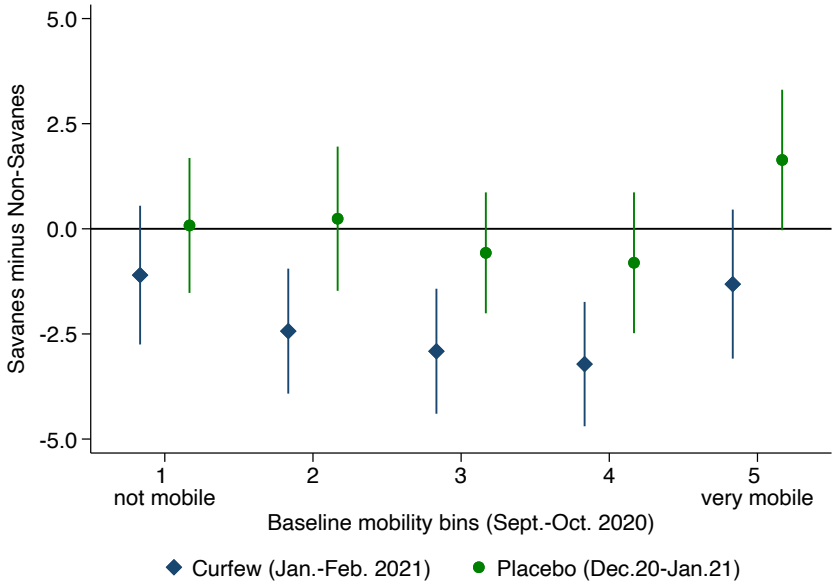

Figure S5: Impacts of GD-Novissi on the aggregated index by baseline mobility quintile